Max Planck Institute for Gravitational Physics

IMPRS Lecture Series

Making sense of data: introduction to statistics for gravitational wave astronomy

Lecturer: Jonathan Gair

Winter, 2019–2020

This course will provide a general introduction to statistics, which will be useful for researchers working in the area of gravitational wave astronomy. It will start with some of the basic ideas from classical (frequentist) and Bayesian statistics then show how some of thee ideas are or will be used in the analysis of data from current and future gravitational wave (GW) detectors. The final section of the course will introduce some advanced topics that are also relevant to GW observations. These topics will not be expounded in great depth, but some of the key ideas will be described to provide familiarity with the concepts. The aim of the course will be to establish sufficient grounding in statistics that students will be able to understand research seminars and papers, and know where to begin if carrying out research in these areas.

The lectures will be supported by a number of computer practicals. Statisticians typically use the community software package R and this is also commonly used by researchers in other disciplines. Most new statistical methods that are developed are implemented as R packages and so familiarity with R will enable the user to carry out fairly sophisticated analyses straightforwardly. However, in physics it is more common these days to use python and there are a number of libraries of statistical functions and methods available for python as well. Therefore, the practicals will use python.

Course outline

1. **(weeks 1–2)** Classical (frequentist) statistics.

   - Random variables: definition, properties, some useful probability distributions, central limit theorem.
   - Statistics: definition, estimators, likelihood, desirable properties of estimators, Cramer-Rao bound.
   - Hypothesis testing: definition, Neyman-Pearson lemma, power and size of tests, type I and type II errors, ROC curves, confidence regions, uniformly-most-powerful tests.

2. **(weeks 3–4)** Bayesian statistics.

   - Bayes' theorem, conjugate priors, Jeffrey's prior.
   - Bayesian hypothesis testing, hierarchical models, posterior predictive checks.
   - Sampling methods for Bayesian inference.

3. **(weeks 5–6)** Statistics in gravitational wave astronomy.

   - Stochastic processes, optimal filtering, signal-to-noise ratio, sensitivity curves.
   - Frequentist statistics in GW astronomy: false alarm rates, Fisher Matrix, PSD estimation.
   - Bayesian statistics in GW astronomy: parameter estimation, population inference, model selection.

4. **(weeks 7–8)** Advanced topics in statistics.

   - Time series analysis: auto-regressive processes, moving average processes, ARMA models.
   - Nonparametric regression: kernel density estimation, smoothing splines, wavelets.
   - Gaussian processes, Dirichlet processes.

1 Random variables

In classical physics most things are deterministic. There are physical laws governing the evolution of a system which can be solved and used to predict the state of the system in the future. In reality there are many situations in which things are not (or effectively not) deterministic, and so the outcome of an experiment cannot be predicted with certainty. However, if the experiment is repeated many times some outcomes will occur more frequently than others. This notion of in-deterministicity in measurements is encoded in the concept of a *random variable*. A random variable, X, is a quantity that, when observed, can take one of a (possibly infinite) number of values. Prior to making a measurement the value of the random variable cannot be predicted, but the relative frequency of the outcomes over many experiments are described by a *probability distribution*. The value that X takes in a particular observation (or experiment), x_i, say is called a *realisation* of the random variable.

Random variables can be *discrete*, in which case the values that the variable takes are drawn from a countable set of discrete possibilities, or *continuous* in which case the random variable may take on any value within one or more ranges.

1.1 Discrete random variables

A discrete random variable X can take on any of a (possibly infinite but countable) set of possible values, $\{x_1, x_2, \ldots)\}$, which together comprise the *sample space*. The probability that X takes any particular value is represented by a *probability mass function* (pmf), which is a set of numbers $\{p_i\}$ with the properties $0 \leq p_i \leq 1$ for all i and $\sum p_i = 1$. The probability that X takes the value x_i is p_i.

1.2 Examples of discrete random variables

1.2.1 Binomial and related distributions

The Binomial distribution is the distribution of the number of success in n trials for which the probability of success in one trial is p. We write $X \sim B(n, p)$ and

$$P(X = k) = p_k = \begin{cases} \left(\begin{array}{c} n \\ k \end{array}\right) p^k (1-p)^{n-k} & \text{if } k \in \{1, \ldots, n\}, \\ 0 & \text{otherwise} \end{cases} . \tag{1}$$

When $n = 1$ this is the *Bernoulli distribution*. The binomial distribution is the distribution of the sum of n Bernoulli trials, i.e., the number of "successes" in n trials. A related distribution is the *negative binomial distribution* which has pmf

$$P(X = k) = p_k = \begin{cases} \left(\begin{array}{c} k+r-1 \\ k \end{array}\right) p^k (1-p)^{r} & \text{if } k \in \{0, 1, \ldots\}, \\ 0 & \text{otherwise} \end{cases} . \tag{2}$$

This is the distribution of the number of successes in a sequence of Bernoulli trials that will be observed before r failures have been observed. Setting $r = 1$ and $p \to (1 - p)$ this is the *geometric distribution*, which is the distribution of the number of trials required before the first success.

Another generalisation of the Binomial distribution is the *multinomial distribution*. In this case the outcome of a trial is not a binary 'success' or 'fail', but it is one of k possible outcomes. The probability of each outcome is denoted p_i with $\sum_{i=1}^k p_i = 1$ and the multinomial distribution describes the probability of seeing n_1 occurrences of outcome 1, n_2 occurrences of outcome 2 etc. in n trials. The pmf is

$$P(\{n1,\ldots,n_k\}) = \left\{ \begin{array}{ll} \frac{n!}{n_1!n_2!\ldots n_k!} p_1^{n_1} p_2^{n_2}\ldots p_k^{n_k} & \text{if } n_i \geq 0 \ \forall i \text{ and } \sum_{i=1}^k n_i = n \\ 0 & \text{otherwise} \end{array} \right. . \tag{3}$$

Applications: counting problems, e.g., distribution of events in categories or time, trials factors.

1.2.2 Poisson distribution

This is the distribution of the number of occurrences of some event in a certain time interval if that event occurs at a *rate* λ. The quantity X follows a Poisson distribution, $X \sim P(\lambda)$ if

$$P(X=k) = p_k = \left\{ \begin{array}{ll} \lambda^k \mathrm{e}^{-\lambda}/k! & \text{if } k \in \{0,1,\ldots\}, \\ 0 & \text{otherwise} \end{array} \right. . \tag{4}$$

The Poisson distribution is the limiting distribution of $B(n,p)$ as $n \to \infty$, $p \to 0$ with $np = \lambda$ fixed.

Applications: distribution of number of events in a population, e.g., gravitational wave sources.

1.3 Continuous random variables

A continuous random variable can take any (usually real, but the extension to complex RVs is straightforward) value within some continuous range, or some set of ranges, which together comprise the *sample space* $\mathcal{X}$. The probability that X takes a particular value is characterised by the *probability density function* (pdf), $p(x)$. The probability that X takes a value in the range x to $x+\mathrm{d}x$ is $p(x)\mathrm{d}x$. The pdf has the properties $0 \leq p(x) \leq 1$ for all $x \in \mathcal{X}$ and

$$\int_{x\in\mathcal{X}} p(x)\mathrm{d}x = 1. \tag{5}$$

For single valued random variables with non-disjoint sample spaces continuous random variables may also be characterised by the *cumulative density function* or CDF, defined as

$$P(X \leq x) = \int_{-\infty}^{x} p(x)\mathrm{d}x. \tag{6}$$

1.3.1 Uniform distribution

X is uniform on an interval (a,b), denoted $X \sim U[a,b]$ if the pdf is constant on the interval $[a,b]$

$$p(x) = \left\{ \begin{array}{ll} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{array} \right. . \tag{7}$$

X takes values only in the range $[a,b]$.

Applications: often used as an "uninformative" prior in parameter estimation.

1.3.2 Normal distribution

X is Normal with *mean* μ and *variance* σ^2, denoted $X \sim N(\mu, \sigma^2)$ if the pdf has the form

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{8}$$

X takes all values in the range $(-\infty, \infty)$. If $\mu = 0$ and $\sigma^2 = 1$ we say that X follows a *standard Normal distribution.*

Applications: distribution of noise fluctuations in a gravitational wave detector, priors on mass distribution, most common distribution to assume in parametric statistics.

1.3.3 Chi-squared distribution

X is chi-squared with k degrees of freedom, denoted $X \sim \chi^2(k)$ or χ^2_k is the pdf has the form

$$p(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{\frac{k}{2}-1} \mathrm{e}^{-\frac{x}{2}} \tag{9}$$

Here $\Gamma(n)$ is the Gamma function, defined by

$$\Gamma(n) = \int_0^\infty x^{n-1}\mathrm{e}^{-x}\mathrm{d}x \tag{10}$$

and such that $\Gamma(n+1) = n!$. X takes non-negative real values only, $x \in [0, \infty)$. This is the distribution of the sum of the squares of n independent standard normal distributions.

There is also a *non-central chi-square distribution* which depends on two parameters — degrees of freedom, $k > 0$, as before plus a *non-centrality parameter*, $\lambda > 0$. This has the pdf

$$p(x) = \frac{1}{2}\mathrm{e}^{-\frac{(x+\lambda)}{2}} \left(\frac{x}{\lambda}\right)^{\frac{k}{4}-\frac{1}{2}} I_{\frac{k}{2}-1}(\sqrt{\lambda x}) \tag{11}$$

where $I_\nu(y)$ is the modified Bessel function of the first kind. The non-central chi-square distribution again takes non-negative values only and arises as the distribution of the sum of k independent normal distributions with equal (unit) variance, but non-zero means, denoted μ_i. The non-centrality parameter is then $\lambda = \sum_{i=1}^k \mu_i^2$.

Applications: used to test for deviations from normality, e.g., in noise fluctuations in a gravitational wave detector.

1.3.4 Student's t-distribution

X follows Student's t-distribution with $n > 0$ degrees of freedom, $X \sim t_n$, if it has pdf

$$p(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}. \tag{12}$$

The Student t-distribution arises in hypothesis testing as the distribution of the ratio of a standard Normal distribution to the square root of an independent χ^2_n distribution, normalised by the degrees of freedom. Specifically if $X \sim N(0, 1$ and $Y \sim \chi^2_n$ then $X/\sqrt{Y/n}$ follows a t_n distribution.

Applications: used for statistical test on significance of parameters in linear models, used as a "heavy-tailed" distribution for robust parameter estimation, arises naturally when marginalising over uncertainty in power-spectral density estimation.

1.3.5 F-distribution

X follows an F-distribution with degrees of freedom $n_1 > 0$ and $n_2 > 0$ if it has pdf

$$p(x) = \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1} \left(1 + \frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}} \tag{13}$$

where $B(a, b)$ is the beta function, which is given by

$$B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}\mathrm{d}x \tag{14}$$

and is related to the Gamma function through $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$. The F-distribution arises as the ratio of two independent chi-squared distributions with n_2 and n_2 degrees of freedom.

Applications: arises primarily in analysis of variance to test differences between groups.

1.3.6 Exponential distribution

X is exponential with *rate* $\lambda > 0$, $X \sim \mathcal{E}(\lambda)$ if it has pdf

$$p(x) = \begin{cases} \lambda\mathrm{e}^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

X takes positive real values only, $x \in (0, \infty)$. The exponential distribution is the distribution of the time that elapses between successive events of a Poisson process.

Applications: distribution of time lag between events, e.g., gravitational wave signals.

1.3.7 Gamma distribution

X is Gamma with parameters $n > 0$ and $\lambda > 0$, $X \sim$Gamma(n, λ), if it has pdf

$$p(x) = \begin{cases} \frac{1}{\Gamma(n)}\lambda^n x^{n-1} e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

X takes positive real values only, $x \in (0, \infty)$. The Gamma distribution is the distribution of of the sum of n exponential distributions with parameter λ.

Applications: conjugate distribution to the Poisson distribution, so useful in Bayesian analysis of rates. Useful as prior distribution whenever variable has support on $[0, \infty)$.

1.3.8 Beta distribution

X is Beta with parameters $a > 0$ and $b > 0$, $X \sim$Beta(a, b), if it has pdf

$$p(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

X takes values in the range $x \in (0, 1)$ only.

Applications: conjugate to binomial distribution. Useful as prior when variable has support on $[0, 1]$, e.g., for probabilities.

1.3.9 Dirichlet distribution

The Dirichlet distribution is a multivariate extension of the Beta distribution. A realisation of a Dirichlet random variable is a set of K values, $\{x_i\}$, satisfying the constraints $0 < x_i < 1$ for all i and $\sum_{i=1}^{K} x_i = 1$. The Dirichlet distribution is characterised by a vector of *concentration parameters* $\vec{\alpha} = (\alpha_1, \ldots, \alpha_K)$ satisfying $\alpha_i > 0$ for all i and has pdf

$$p(x) = \frac{1}{B(\vec{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \qquad \text{where } B(\vec{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{j=1}^{K} \alpha_j\right)}. \tag{18}$$

Applications: infinite dimensional generalisation is a Dirichlet process which is used as a distribution on probability distributions. Very important in Bayesian nonparametric analysis.

1.3.10 Cauchy distribution

X follows a Cauchy distribution (also known as a Lorentz distribution) with *location parameter* x_0 and *scale parameter* $\gamma > 0$, if it has pdf

$$p(x) = \frac{1}{\pi\gamma \left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]}. \tag{19}$$

X takes any real value $x \in (-\infty, \infty)$. The Cauchy distribution arises as the distribution of the x intercept of a ray issuing from the point (x_0, γ) with a uniformly distributed angle. It is also the distribution of the ratio of two independent zero-mean Normal distributions.

Applications: used to model distributions with sharp features. In a gravitational wave context it is used as a model for lines in the spectral density of gravitational wave detectors, for example in BayesLine (and hence BayesWave).

1.4 Properties of random variables

The pdf (or pmf) of a random variable tells us everything about the random variable. However, it is often convenient to work with a smaller number of quantities that summarise the properties of the distribution. These characterise the ‘average’ value of a random variable and the spread of the random variable about the average. We summarise a few of these quantities here. They all rely on the notion of an *expectation value*, denoted $\mathbb{E}$. The expectation value of a function, $T(X)$, of a discrete random variable X is defined by

$$\mathbb{E}(T(X)) = \sum_{i=1}^{\infty} p_i t(x_i). \tag{20}$$

A similar definition holds for continuous random variables by replacing the sum with an integral

$$\mathbb{E}(T(X)) = \int_{-\infty}^{\infty} p(x) t(x) \mathrm{d}x. \tag{21}$$

1.4.1 Quantities representing the average value of a random variable

- **Mean** The mean, often denoted μ, is the expectation value of X, $\mu = \mathbb{E}(X)$.

- **Median** The median, m, is the central value of the distribution in probability, i.e., a value such that the probability of obtaining a value smaller than that or larger than that is (roughly) equal. For discrete random variables $m = x_k$, where

$$\sum_{i:x_i<x_k} p_i < 0.5 \qquad \text{and} \sum_{i:x_i\leq x_k} p_i \geq 0.5. \tag{22}$$

  For continuous random variables m is the value such that

$$\int_{-\infty}^{m} p(x)\mathrm{d}x = \int_{m}^{\infty} p(x)\mathrm{d}x = \frac{1}{2}. \tag{23}$$

- **Mode** The mode, M, is the 'most probable' value of the random variable. For discrete random variables

$$M = \mathrm{argmax}_{i\in\mathcal{X}} p_i \tag{24}$$

  and for continuous random variables

$$M = \mathrm{argmax}_{x\in\mathcal{X}} p(x). \tag{25}$$

  The mode may not be unique.

1.4.2 Quantities representing the spread of a random variable

- **Variance** The variance, often denoted σ^2, is the expectation value of the squared distance from the mean, i.e.,

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right]. \tag{26}$$

- **Standard deviation** The standard deviation is simply the square root of the variance, usually denoted σ.

- **Covariance** When considering two random variables, X and Y say, the covariance is defined as the expectation value of the product of their distance from their respective means, i.e.,

$$\mathrm{cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right]. \tag{27}$$

  Here the expectation value is taken with respect to the joint distribution (see section on independence below).

- **Skewness** Given the mean, μ, and variance, σ^2, defined above, the skewness of a distribution is

$$\gamma_1 = \mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^3\right]. \tag{28}$$

- **Kurtosis** In a similar way, kurtosis is defined as

$$\mathrm{Kurt}(X) = \mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^4\right]. \tag{29}$$

  This measures the heaviness of the tails of the distribution of the random variable. The kurtosis of the Normal distribution is 3, so it is common to quote *excess kurtosis*, which is the kurtosis minus 3, i.e., the excess relative to the Normal distribution.

- **Higher moments** Higher moments can be defined in a similar way. The n'th moment about a reference value c of a probability distribution is

$$\mathbb{E}\left[(X-c)^n\right]. \tag{30}$$

  Moments are usually defined with c taken to be the mean, μ, as in the definition of skewness and kurtosis above.

1.4.3 Moment generating functions

A useful object for computing summary quantities of a probability distribution is the *moment generating function*, $M_X(t)$, which is defined as

$$M_X(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right] \quad t \in \mathbb{R}. \tag{31}$$

It is clear that derivatives of this function with respect to t, evaluated at $t = 0$, give successive moments about zero of the distribution. Moment generating functions (MGFs) are defined in the same way for both discrete and continuous random variables.

In Table 1 we list these various summary quantities for the probability distributions listed earlier. Where quantities are not known in closed form they are omitted from this table.

Distribution	Mean	Median	Mode	Variance	Skewness	Excess kurtosis	MGF
Binomial(n,p)	np	$\lfloor np \rfloor$	$\lfloor (n+1)p \rfloor$	$np(1-p)$	$\frac{1-2p}{\sqrt{np(1-p)}}$	$\frac{1-6p(1-p)}{np(1-p)}$	$(1-p+p\mathrm{e}^t)^n$
Poisson(λ)	λ	$\approx \lfloor \lambda + \frac{1}{3} - \frac{0.02}{\lambda} \rfloor$	$\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$	λ	$\lambda^{-\frac{1}{2}}$	λ^{-1}	$\exp\left[\lambda(\mathrm{e}^t - 1)\right]$
Uniform$[a,b]$	$\frac{1}{2}(a+b)$	$\frac{1}{2}(a+b)$	all	$\frac{1}{12}(b-a)^2$	0	$-\frac{6}{5}$	$\frac{\mathrm{e}^{tb}-\mathrm{e}^{ta}}{t(b-a)}$
Normal(μ,σ^2)	μ	μ	μ	σ^2	0	0	$\exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]$
χ^2_n	n	$\approx n\left(1-\frac{2}{9n}\right)^3$	$\max(n-2,0)$	$2n$	$\sqrt{\frac{8}{n}}$	$\frac{12}{n}$	$(1-2t)^{-k/2}$
Student's t_n	0	0	0	$\frac{n}{n-2}$	0 for $n > 3$	$\frac{6}{n-4}$ for $n > 4$	—
F(n_1,n_2)	$\frac{n_1}{n_2-2}$	—	$\frac{n_2(n_1-2)}{n_1(n_2+2)}$	$\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$	$\frac{(2n_1+n_2-2)\sqrt{8(n_2-4)}}{(n_2-6)\sqrt{n_1(n_1+n_2-2)}}$	see caption	—
$\mathcal{E}(\lambda)$	$\frac{1}{\lambda}$	$\frac{\ln 2}{\lambda}$	0	$\frac{1}{\lambda^2}$	2	6	$\frac{\lambda}{\lambda-t}$
Gamma(n,λ)	$\frac{n}{\lambda}$	—	$\frac{n-1}{\lambda}$	$\frac{n}{\lambda^2}$	$\frac{2}{\sqrt{n}}$	$\frac{6}{n}$	$\left(1-\frac{t}{\lambda}\right)^{-n}$
Beta(a,b)	$\frac{a}{a+b}$	$I^{[-1]}_{\frac{1}{2}}(a,b)$	$\frac{a-1}{a+b-2}$	$\frac{ab}{(a+b)^2(a+b+1)}$	$\frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}$	see caption	see caption
Dirichlet $(K,\vec{\alpha})$	$\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}$	—	$\frac{\alpha_i-1}{\sum_{j=1}^K \alpha_j - K}$	$\frac{\bar{\alpha}_i(1-\bar{\alpha}_i)}{\alpha_0+1}$	—	—	—
Cauchy (x_0,γ)	undefined	x_0	x_0	undefined	undefined	undefined	does not exist

Table 1: Summary of important properties of common probability distributions. The excess kurtosis of the F distribution is $12n_1(5n_2-22)(n_1+n_2-2)+(n_2-4)(n_2-2)^2/[n_1(n_2-6)(n_2-8)(n_1+n_2-2)]$. For the Beta$(a,b)$ distribution, the excess kurtosis is $6[(a-b)^2(a+b+1)-ab(a+b+2)]/[ab(a+b+2)(a+b+3)]$ and the MGF is $1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{a+r}{a+b+r}\right)\frac{t^k}{k!}$. For the Dirichlet distribution, the mean and variance are quoted for one component of the distribution, x_i, the parameters $\alpha_0 = \sum_{j=1}^K \alpha_j$ and $\bar{\alpha}_i = \alpha_i / \sum_{j=1}^K \alpha_j$ and the covariance $\mathrm{cov}(x_i,x_j) = -\bar{\alpha}_i\bar{\alpha}_j/(1+\alpha_0)$.

1.5 Independence

Most of the random variables described above are single valued, but a few of them, e.g., the multinomial and Dirichlet distributions, return multiple values. In other situations, several random variables might be evaluated simultaneously, or sequentially, or the same random variable might be observed multiple times. When dealing with multiple random variables, covariance as introduced above is an important concept, as is *independence*. A set of random variables $\{X_1, \ldots, X_N\}$ are said to be *independent* if

$$P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_N \leq x_N) = P(X_1 \leq x_1)P(X_1 \leq x_1)\ldots P(X_1 \leq x_1) \quad \forall\, x_1,\, x_2,\, \ldots,\, x_N. \tag{32}$$

In terms of the pdf (or pmf) the random variables are independent if their joint distribution $p(x_1, \ldots, x_N)$ can be separated

$$p(x_1, \ldots, x_N) = p_{X_1}(x_1)p_{X_2}(x_2)\ldots p_{X_N}(x_N). \tag{33}$$

Independence of two random variables implies that the covariance is 0, but the converse is not true except in certain special cases, for example for two Normal random variables.

A set of variables $\{X_i\}$ is called *independent identically distributed* or IID if they are independent and all have the same probability distribution. This situation arises often, for example when taking multiple repeated observations with an experiment.

1.6 Linear combinations of random variables

Suppose $X_1, \ldots, X_N$ are (not necessarily independent) random variables and consider a new random variable Y defined as

$$Y = \sum_{i=1}^{N} a_i X_i. \tag{34}$$

For any set of random variables

$$\mathbb{E}(Y) = \sum_{i=1}^{N} a_i \mathbb{E}(X_i), \qquad \mathrm{Var}(Y) = \sum_{i=1}^{N} a_i^2 \mathrm{Var}(X_i) + \sum_{i \neq j} a_i a_j \mathrm{cov}(X_i, X_j). \tag{35}$$

If the random variables are *independent* then the variance expression simplifies to

$$\mathrm{Var}(Y) = \sum_{i=1}^{N} a_i^2 \mathrm{Var}(X_i) \tag{36}$$

and the moment generating function of Y can be found to be

$$M_Y(t) = \prod_{i=1}^{N} M_{X_i}(a_i t). \tag{37}$$

A commonly used linear combination of random variables is the *sample mean* of a set of IID random variables, defined as

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{38}$$

for which

$$\mathbb{E}(\hat{\mu}) = \mathbb{E}(X_1), \qquad \mathrm{Var}(\hat{\mu}) = \frac{1}{n}\mathrm{Var}(X_1), \qquad M_{\hat{\mu}}(t) = \left(M_{X_1}\left(\frac{t}{N}\right)\right)^N. \tag{39}$$

1.7 Laws of large numbers

Suppose that $X_1, \ldots, X_n$ are a sequence of IID random variables, each having finite mean μ and variance σ^2. We denote the sum of the random variables by

$$S_n = \sum_{i=1}^{n} X_i, \qquad \text{which implies } \mathbb{E}(S_n) = n\mu, \quad \mathrm{Var}(S_n) = n\sigma^2. \tag{40}$$

Laws of large numbers tells us that the sample mean becomes increasingly concentrated around the mean of the random variable as the number of samples tends to infinity.

1.7.1 Weak law of large numbers

The *weak law of large numbers* states that, for $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \to 0, \text{ as } n \to \infty. \tag{41}$$

1.7.2 Strong law of large numbers

The *strong law of large numbers* states simply

$$P\left(\frac{S_n}{n} \to \mu\right) = 1. \tag{42}$$

1.7.3 Central limit theorem

In many applications, people assume that the data generating process is Normal. This is partially because the Normal distribution is convenient to work with and has many nice properties, but also because regardless of the distribution large samples of random variables tend to look quite Normally distributed. This fact is encoded in the *Central Limit Theorem*, which states that the standardized sample mean, S_n^*, is approximately standard Normal in the limit $n \to \infty$

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}. \tag{43}$$

Formally the statement of the central limit theorem is

$$\lim_{n\to\infty} P(a \leq S_n^* \leq b) = \Phi(b) - \Phi(a) = \lim_{n\to\infty} P(n\mu + a\sigma\sqrt{n} \leq S_n \leq n\mu + b\sigma\sqrt{n}). \tag{44}$$

2 Frequentist statistics

In the last section we discussed the notion of a random variable. When observing phenomena in nature or performing experiments we would like to deduce the distribution of the random variable, i.e., the probability distribution from which realisations of that random variable are drawn. In **parametric inference** we assume that the distribution of the random variable takes a particular form, i.e., it belongs to a known family of probability distributions. All of the distributions that were described in the previous section are characterised by one or more parameters and so inference about the form of the distribution reduces to inference about the values of those parameters.

In *frequentist statistics* we assume that the parameters characterising the distribution are *fixed* but *unknown*. Statements about the parameters, for example *significance* and *confidence* are statements about multiple repetitions of the same observation, with the parameters fixed. Key frequentist concepts are *statistics*, *estimators* and *likelihood*.

A **statistic** is a random variable or random vector $T = t(\mathbf{X})$ which is a function of $\mathbf{X}$ but does not depend on the parameters of the distribution, θ. Its realised value is $t = t(\mathbf{x})$. In other words a statistic is a function of observed data only, not the unknown parameters.

An **estimator** is a statistic used to estimate the value of a parameter. Typically the random vector would be a set of IID random variables, $X_1, \ldots, X_n$ with pdf $p(x|\,\theta)$. A function $\widehat{\theta}(X_1, \ldots, X_n)$ of $X_1, \ldots, X_n$ used to infer the parameter values is called an **estimator** of θ; note that $\widehat{\theta}$ is a random variable with a sampling distribution in this latter context. The value of the estimator at the observed data $\widehat{\theta}(x_1, \ldots, x_n)$ is called an **estimate** of θ.

A statistic might also be used to provide an upper or lower limit for a *confidence interval* on the value of a parameter, or to evaluate the validity of a hypothesis in *hypothesis testing*.

2.1 Likelihood

Likelihood is central to the theory of frequentist parametric inference.

If an event E has probability which is a specified function of parameters $\vec{\theta}$, then the likelihood of E is $\mathbb{P}(E|\,\vec{\theta})$, regarded as a function of $\vec{\theta}$.

The likelihood, denoted $L(\vec{\theta};\mathbf{x})$, is functionally the same as the pdf of the data generating process, the difference is that the likelihood is regarded as a function of the parameters $\vec{\theta}$ while the pdf is regarded as a function of the observed data, $\mathbf{x}$. It is often convenient to work with the **log likelihood**

$$l(\theta;\mathbf{x}) = \ln[L(\theta;\mathbf{x})] = \ln[p(\mathbf{x}|\,\theta)] \qquad (\theta \in \Theta)$$

Another useful quantity is the **score**

$$\frac{\partial l}{\partial \theta_i}$$

which is a vector that is also regarded as a function of $\vec{\theta}$ with the data fixed at the observed values.

One interpretation of likelihood is that, given data $\mathbf{x}$, the relative plausibility of or support for different values $\vec{\theta}_1$, $\vec{\theta}_2$ of $\vec{\theta}$ is expressed by

$$\frac{L(\vec{\theta}_1;\mathbf{x})}{L(\vec{\theta}_2;\mathbf{x})} \qquad \text{or} \qquad l(\vec{\theta}_1;\mathbf{x}) - l(\vec{\theta}_2;|\mathbf{x}).$$

As a result, inferences are unchanged if $L(\vec{\theta}|\mathbf{x})$ is multiplied by a positive constant (possibly depending on $\mathbf{x}$).

Typically we will be interested in cases where we observe more than one independent realisation of the random variable. For discrete random variables the combined likelihood is then the product of the likelihoods of each observed event.

Example: Poisson distribution

We observe a set $\{x_1, \ldots, x_n\}$, of n IID observations from a Poisson distribution with parameter λ. Denoting $n\bar{x} = \sum_{j=1}^{n} x_j$ the likelihood is

$$L(\theta; \mathbf{x}) = \frac{e^{-n\lambda}\lambda^{n\bar{x}}}{\prod_j x_j!} \quad (\lambda > 0)$$

$$l(\lambda; \mathbf{x}) = \log\left(L(\lambda; \mathbf{x})\right] = -n\lambda + n\bar{x}\ln\lambda - \ln(\prod_j x_j!)$$

For continuous random variables the joint likelihood can usually be written as

$$L(\theta; \mathbf{x}) = \prod_{j=1}^{n} p(x_j|\,\theta) \qquad \Rightarrow \qquad l(\theta; \mathbf{x}) = \sum_{j=1}^{n} l(x_j|\,\theta).$$

or just $p(\mathbf{x}|\,\theta)$ for a vector $\mathbf{x}$ of random variables that are not IID. One case where this does not necessarily hold is when measurements are imperfect. Typically we cannot observe a quantity with infinite precision, but inevitably round to the nearest measurement unit. Observations of continuous random variables therefore typically involve grouping measurements into bins.

Suppose random variables $X_1, \ldots, X_n$ are IID with cumulative distribution function $P(x|\,\vec{\theta})$ and we observe that there are $n_1, \ldots, n_k$ observations in each of the k intervals $(a_0, a_1], \ldots, (a_{k-1}, a_k]$, where $-\infty \leq a_0 < a_1 < \ldots < a_k \leq \infty$ and $\mathbb{P}(a_0 < X_j \leq a_k) = 1$. The distribution of $(N_1, \ldots N_k)$ is Multinomial with parameters $(n, p_1(\vec{\theta}), \ldots p_k(\vec{\theta}))$ with

$$p_r(\vec{\theta}) = \mathbb{P}(a_{r-1} < X_j \leq a_r|\,\vec{\theta}) = P(a_r|\,\vec{\theta}) - P(a_{r-1}|\,\vec{\theta}),$$

and the likelihood is given by (3). For example, with common distribution $N(\mu, \sigma^2)$ we have

$$p_r(\mu, \sigma^2) = \Phi\left(\frac{a_r - \mu}{\sigma}\right) - \Phi\left(\frac{a_{r-1} - \mu}{\sigma}\right).$$

If observations of the IID random variables are made with a resolution (or maximum grouping error)of $\pm\frac{1}{2}h$, then we are effectively in the above situation, and a recorded value x represents a value in the range $x \pm \frac{1}{2}h$. Assuming that the grouping error is small, the likelihood is

$$\prod_{j=1}^{n}\{P(x_j + \frac{1}{2}h|\,\theta) - P(x_j - \frac{1}{2}h|\,\theta)\}. \tag{45}$$

If $p(x|\,\theta)$ does not vary too rapidly in each interval $(x_j - \frac{1}{2}h, x_j + \frac{1}{2}h)$ then (45) can be approximated by

$$\prod_{j=1}^{n}\{hp(x_j|\,\theta)\},$$

or, ignoring the constant h^n,

$$L(\theta;\mathbf{x}) \simeq \prod_{j=1}^{n} p(x_j|\,\theta).$$

which is the result we wrote down when there was no grouping error. However, this argument can fail, as illustrated in the two examples below.

Examples where this approximation fails

- Single observation from $N(\mu,\sigma^2)$

$$\begin{aligned} L(\mu,\sigma|x) &= \Phi\left\{\frac{x+\frac{1}{2}h-\mu}{\sigma}\right\} - \Phi\left\{\frac{x-\frac{1}{2}h-\mu}{\sigma}\right\} && (46)\\ &\simeq \frac{h\ \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)}{\sqrt{2\pi}\sigma} && (47)\end{aligned}$$

  if $\sigma > h$. If $\mu = x$ and $\sigma \to 0$, (46)$\to 1$ but (47)$\to \infty$.

- Uniform distribution on $[0,\theta]$, $U(0,\theta)$
  If $X_1,\ \ldots\ ,X_n$ are IID with pdf given by

$$p(x|\,\theta) = \begin{cases} \frac{1}{\theta} & (0 < x \le \theta)\\ 0 & \text{otherwise}\end{cases}$$

  then

$$p(\mathbf{x}|\,\theta) = \begin{cases} \frac{1}{\theta^n} & (0 < x_{(n)} \le \theta)\\ 0 & \text{otherwise}\end{cases}$$

  where $x_(i)$ denotes the i'th element in the ordered sequence of $\{x_i\}$. The likelihood is

$$L(\theta;\mathbf{x}) \simeq \begin{cases} 0 & (\theta < x_{(n)})\\ \frac{1}{\theta^n} & (\theta \ge x_{(n)})\end{cases} \qquad (48)$$

  Taking account of a grouping error of $\pm\frac{1}{2}h$, the probability assigned to $(x_j-\frac{1}{2}h, x_j+\frac{1}{2}h)$ is

$$\begin{cases} \frac{h}{\theta} & (x_j+\frac{1}{2}h < \theta)\\ \frac{\theta - x_j + \frac{1}{2}h}{\theta} & (x_j - \frac{1}{2}h \le \theta < x_j + \frac{1}{2}h)\end{cases}$$

  and, if $h \le x_{(n)} - x_{(n-1)}$,

$$L(\theta;\mathbf{x}) \propto \begin{cases} 0 & (\theta < x_{(n)} - \frac{1}{2}h)\\ \frac{\left[(\theta - x_{(n)} + \frac{1}{2}h)/h\right]^a}{\theta^n} & (x_{(n)} - \frac{1}{2}h \le \theta < x_{(n)} + \frac{1}{2}h)\\ \frac{1}{\theta^n} & (\theta > x_{(n)} + \frac{1}{2}h)\end{cases} \qquad (49)$$

  where a is the number of observations equal to $x_{(n)}$. The continuous likelihood (Eq. (48)) and the likelihood accounting for grouping error (Eq. (49)) are shown in Figure 1.

  Ignoring grouping, $x_{(n)}$ is the ML estimator and has variance of order n^{-2}; with grouping the asymptotic variance is the usual $O(n^{-1})$.

To summarise: if the precision of observing the data (h) is much smaller than the variability of the data (e.g. than the standard deviation) then it is fine to use the approximation of the likelihood by the density. However, if the precision h is comparable with the variability, in order to estimate the unknown parameters reliably, one has to use the discrete version of the likelihood.

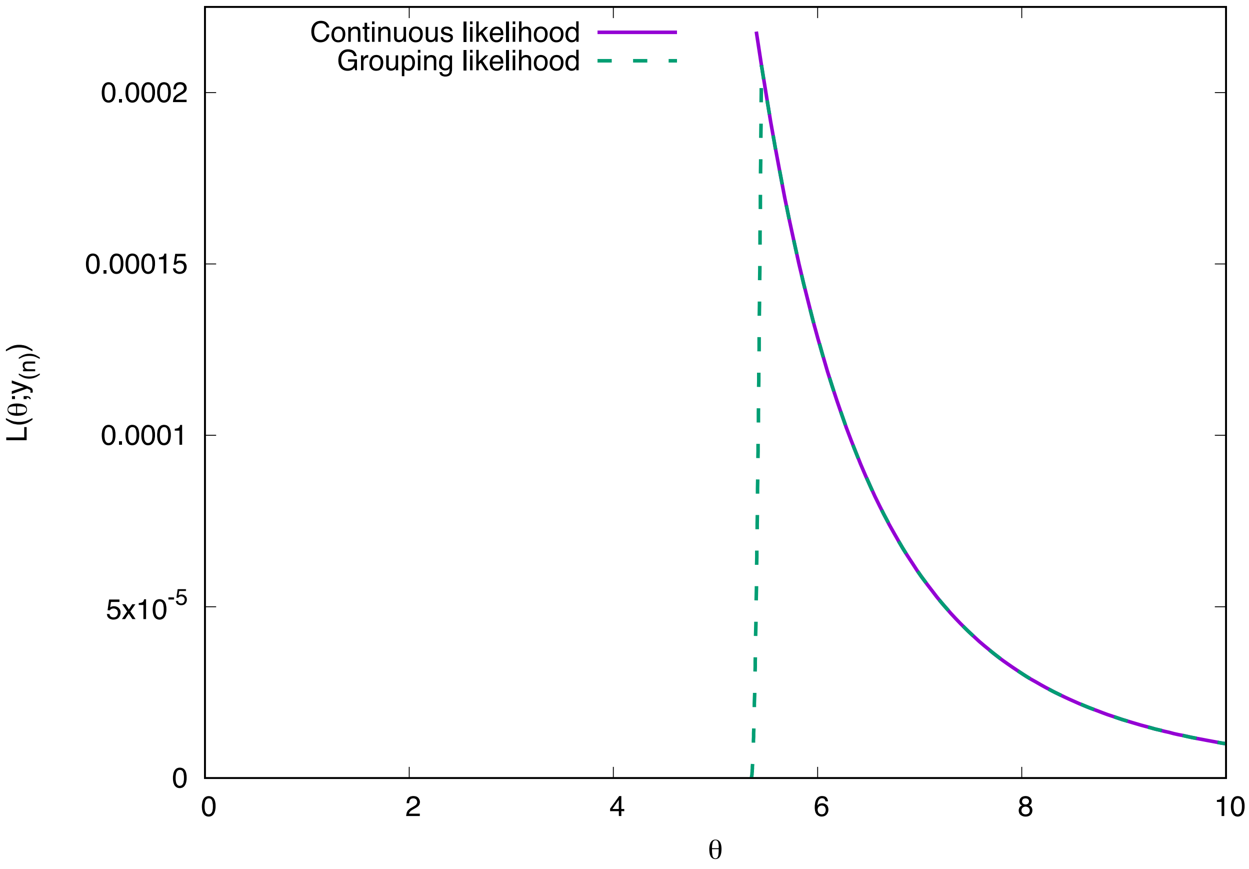


Figure 1: The continuous likelihood for the parameter, θ, of the uniform distribution, as given in Eq. (48), based on $n = 5$ observations with maximum observed value $x_{(n)} = 5.4$ (solid purple line). Also shown is the likelihood including grouping error, as given in Eq. (49), assuming that results are rounded to one decimal place, $h = 0.1$, and there are $a = 2$ observations equal to 5.4 (dashed green line).

2.2 Sufficient statistics

If a parametric form is assumed for the distribution of X, then there may exist a lower dimensional function of the vector of observations $\mathbf{x}$ that contains the same information on the value of $\vec{\theta}$ as vector $\mathbf{x}$. Such a function is called a **sufficient statistic**.

2.3 Definition

Suppose a random vector $\mathbf{X}$ has distribution function in a parametric family $\{P(\mathbf{x}|\,\theta); \theta \in \Theta\}$ and realized value $\mathbf{x}$. A statistic (recall this just means a function of observed data only) is said to be **sufficient** for $\vec{\theta}$ if the distribution of $\mathbf{X}$ given S does not depend on $\vec{\theta}$, i.e. $p_{\mathbf{X}|S}(\mathbf{X}|s, \vec{\theta})$ does not depend on $\vec{\theta}$. Note that

(i) if S is sufficient for $\vec{\theta}$, so is any one-to-one function of S.

(ii) $\mathbf{X}$ is trivially sufficient.

Examples

- Bernoulli trials : $X_1, \ \ldots \ , X_n$ take values 0 or 1 independently with probabilities $1-p$ and p; n is fixed.

$$p_{\mathbf{X}}(\mathbf{x}|\,p) = \prod_{j=1}^{n} p^{x_j}(1-p)^{1-x_j} = p^{\sum x_j}(1-p)^{n-\sum x_j} \tag{50}$$

If $S = X_1 + \ \cdots \ + X_n$, then S has the Binomial p.d.f.

$$p_S(s|\,p) = \left(\begin{array}{c} n \\ s \end{array} \right) p^s (1-p)^{n-s} \quad (s = 0, 1, \ \ldots \ , n)$$

and the p.d.f. of $\mathbf{X}$ given S is

$$\begin{aligned} p_{\mathbf{X}|s}(\mathbf{x}|s) &= \frac{\mathbb{P}(X_1 = x_1, \ \ldots \ , X_n = x_n, X_1 + \ \cdots \ + X_n = s|\,\theta)}{\mathbb{P}(X_1 + \ \cdots \ + X_n = s)} \\ &= \begin{cases} \frac{p_{\mathbf{X}}(\mathbf{x}|\,p)}{p_S(s|\,p)} & (\sum x_j = s) \\ 0 & (\sum x_j \neq s) \end{cases} \\ &= \begin{cases} \left(\begin{array}{c} n \\ s \end{array} \right)^{-1} & (\sum x_j = s) \\ 0 & (\sum x_j \neq s) \end{cases} \end{aligned}$$

This does not depend on p, so S is sufficient for p.

For example, in the case when $n = 3$ the conditional p.d.f of $\mathbf{x} = (x_1, x_2, x_3)$ given $s = \sum x_i$ is as follows:

Sample	$s = \sum x_i$			
(y_1, y_2, y_3)	0	1	2	3
(0 0 0)	1	0	0	0
(1 0 0)	0	$\frac{1}{3}$	0	0
(0 1 0)	0	$\frac{1}{3}$	0	0
(0 0 1)	0	$\frac{1}{3}$	0	0
(1 1 0)	0	0	$\frac{1}{3}$	0
(1 0 1)	0	0	$\frac{1}{3}$	0
(0 1 1)	0	0	$\frac{1}{3}$	0
(1 1 1)	0	0	0	1

- Pois(λ) , $S = X_1 + \cdots + X_n$ has distribution Pois($n\lambda$) and p.d.f.

$$p_S(s|\,\lambda) = \frac{e^{-n\lambda}(n\lambda)^s}{s!},$$

so the distribution of $\mathbf{X}$ given s has p.d.f.

$$p_{X|s}(X|s) = \begin{cases} \frac{p_{\mathbf{X}}(\mathbf{x}|\,\lambda)}{p_S(s|\,\lambda)} = \frac{e^{-n\lambda}\lambda^{\sum x_j}(\prod_j x_j!)^{-1}}{\frac{e^{-n\lambda}(n\lambda)^s}{s!}} = \frac{n^{-s}s!}{\prod_j x_j!} & (\sum x_j = s) \\ 0 & (\sum x_j \neq s) \end{cases},$$

which does not depend on λ (it is a multinomial distribution), so S is sufficient for λ.

Interpretation of sufficiency: If S is sufficient for $\vec{\theta}$, we can argue that $\mathbf{x}$ contains no information on $\vec{\theta}$ beyond what is contained in the value s of S, i.e. all the information in $\mathbf{X}$ about $\vec{\theta}$ is contained in s. This suggests that inferences about the value of $\vec{\theta}$ should be based on the value of s. The rest of the information in $\mathbf{y}$ is still relevant to testing the correctness of the assumed parametric family, e.g., by a residual analysis. Sufficiency leads to replacing $\mathbf{x}$ by s and hence to a reduction in the data, so there is an advantage in using statistical models and designs which lead to sufficient statistics of low dimensionality.

2.4 Recognizing sufficient statistics: Neyman Factorization Theorem

Theorem 2.1. *(Neyman Factorization Theorem). Let $\mathbf{X} = (X_1, \ldots, X_n) \sim p(\mathbf{x}|\,\vec{\theta})$. Then, statistic $s = s(X_1, \ldots, X_n)$ is sufficient for θ iff there exist functions h of $\mathbf{x}$ and g of $(s, \vec{\theta})$ such that*

$$p(\mathbf{x} \mid \vec{\theta}) = L(\vec{\theta}; \mathbf{x}) = g(s(\mathbf{x}), \vec{\theta})h(\mathbf{x}) \quad \forall \vec{\theta} \in \Theta, \ \mathbf{x} \in \mathcal{X} \tag{51}$$

Proof. Proof (discrete case only).

If s is sufficient, then the conditional p.d.f. $p_{\mathbf{X}|S}(\mathbf{x}|s)$ does not depend on $\vec{\theta}$ and we can take $h(\mathbf{x})$ to be $p_{\mathbf{X}|S}(\mathbf{x}|s)$ and $g(s;\theta)$ to be $f_S(s|\,\theta)$. Then

$$\begin{aligned} L(\vec{\theta}; \mathbf{x}) = p_{\mathbf{X}}(\mathbf{x}|\,\vec{\theta}) &= \mathbb{P}(\mathbf{X} = \mathbf{x}|\,\vec{\theta}) \\ &= \mathbb{P}(\mathbf{X} = \mathbf{x} \,\&\, S = s(\mathbf{x}) \mid \vec{\theta}) \\ &= \mathbb{P}(\mathbf{X} = \mathbf{x}|S = s(\mathbf{x}), \vec{\theta})\, \mathbb{P}(S = s(\mathbf{x})|\,\vec{\theta}) \\ &= \mathbb{P}(\mathbf{X} = \mathbf{x}|S = s(\mathbf{x}))\, \mathbb{P}(S = s(\mathbf{x})|\,\vec{\theta}) \ \text{[since } S \text{ is sufficient]} \\ &= h(\mathbf{x})g(s(\mathbf{x}), \vec{\theta}). \end{aligned}$$

Conversely, if (51) holds, then for any given s there is a subset A_s of $\mathcal{X}$ in which $s(\mathbf{x}) = s$; for $\mathbf{x}$ in A_s

$$\mathbb{P}(\mathbf{X} = \mathbf{x}|S = s, \vec{\theta}) = \frac{f_{\mathbf{X}}(\mathbf{x}|\,\vec{\theta})}{\sum_{\mathbf{z}\in A_s} f_{\mathbf{X}}(\mathbf{z}|\,\vec{\theta})} = \frac{h(\mathbf{x})}{\sum_{\mathbf{z}\in A_s} h(\mathbf{z})},$$

while for $\mathbf{x} \notin A_s$ $\mathbb{P}(\mathbf{X} = \mathbf{x}|S = s, \vec{\theta}) = 0$. Thus the conditional distribution does not depend on $\vec{\theta}$, i.e. S is sufficient for $\vec{\theta}$.

□

Note: the statistic $s(\mathbf{x})$ divides the sample space $\mathcal{X}$ into equivalence classes A_s (one for each value of s). This partitioning of $\mathcal{X}$ is unchanged if s is replaced by any one-to-one function of s.

Examples

- Bernoulli trials
$$L(p; \mathbf{x}) = p^{\sum x_j}(1-p)^{n-\sum x_j},$$
so if $s(\mathbf{x}) = \sum x_j$, we could take $h(\mathbf{x}) = 1$, $g(s,p) = p^s(1-p)^{n-s}$
[or, alternatively, we could take $h(\mathbf{x}) = \begin{pmatrix} n \\ s \end{pmatrix}^{-1}$, $g(s,p) = \begin{pmatrix} n \\ s \end{pmatrix} p^s(1-p)^{n-s}$].

- Pois(λ), with $s = \sum x_i$ we have the factorization
$$L(\lambda; \mathbf{x}) = (\prod x_j!)^{-1} \cdot e^{-n\lambda}\lambda^s$$

- The Gamma distribution $\Gamma(\alpha, \lambda)$
$$p_{\mathbf{X}}(\mathbf{x}|\,\alpha, \lambda) = \prod_{j=1}^{n}\left[\frac{\lambda^\alpha x_j^{\alpha-1}e^{-\lambda x_j}}{\Gamma(\alpha)}\right] = \frac{\lambda^{n\alpha}(\prod_j x_j)^{\alpha-1}e^{-\lambda\sum x_j}}{\{\Gamma(\alpha)\}^n} = 1 \cdot \frac{\lambda^{n\alpha}(s_2)^{\alpha-1}e^{-\lambda s_1}}{\{\Gamma(\alpha)\}^n}$$
Therefore, $(s_1, s_2) = (\sum x_j, \prod x_j)$ is sufficient for (α, λ).

- In a gravitational wave context, reduced order models are used to form a basis for the space of waveforms. Given a set $\{h_i(t)\}$ of basis functions that describe a waveform model, the set $\{(\mathbf{d}|\mathbf{h}_i)\}$ of overlaps of the basis functions with the data are sufficient statistics for deducing the waveform parameters.

2.5 Minimal sufficiency

(Non-trivial) sufficiency leads to a reduction in the data; sufficient statistics achieving the greatest reduction are called **minimal sufficient**, i.e. a minimal sufficient statistic is a function of all other sufficient statistics.

While such statistics are usually obvious, a general method for finding them is implied from the following lemma.

Lemma 2.1. *Consider the following partition of the sample space of* $\mathbf{X} = (X_1, \ldots, X_n) \in \mathcal{X}^n$*:* $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$ *belong to the same class of the partition if and only if* $L(\vec{\theta}; \mathbf{x})/L(\vec{\theta}; \mathbf{y})$ *does not depend on* $\vec{\theta}$*.*

Then, any statistic defining this partition is minimal sufficient.

Example

- Weibull distribution: $\{X_1, \ \ldots \ , X_n\}$ IID from Weibull with pdf

  $$p(y|\,\alpha, \lambda) = \alpha\lambda^{\alpha} x^{\alpha-1} \exp[-(\lambda x)^{\alpha}] \quad (x > 0; \alpha, \lambda > 0)$$

  Then

  $$L(\alpha, \lambda; \mathbf{x}) = \alpha^n \lambda^{n\alpha} (\prod_{j=1}^{n} x_j)^{\alpha-1} \exp(-\lambda^{\alpha} \sum x_j^{\alpha})$$

  For $L(\alpha, \lambda; \mathbf{z})/L(\alpha, \lambda; \mathbf{x})$ not to depend on α, λ, the z_j must be some permutation of the x_j, but no other reduction in the data retains sufficiency, i.e. the order statistics $x_{(1)} \leq \ \ldots \ \leq x_{(n)}$ are minimal sufficient.

2.6 Exponential families of distributions

A family of distributions indexed by a multivariate parameter $\vec{\theta} \in \Theta \subset \mathbb{R}^p$, is an **exponential family** iff for some real-valued functions $\{A_j; j = 1 \ldots, K\}, \{B_j; j = 1 \ldots, K\}, C, D$ the pdf has the form

$$p(x|\,\theta) = \exp\left\{\sum_{j=1}^{K} A_j(x) B_j(\vec{\theta}) + C(\vec{\theta}) + D(x)\right\} \ \forall x, \vec{\theta} \tag{52}$$

Given observations $\{x_1, \ldots, x_n\}$, the set of K statistics $\{\sum_{j=1}^{n} A_i(x_j) \, : \, 1 \leq i \leq K\}$ are sufficient for $\vec{\theta}$ and they are called the natural statistics of the exponential family

In fact, for a K-dimensional parameter $\vec{\theta}$, the minimal sufficient statistic vector is also K-dimensional only for the distributions from the exponential family (under certain regularity conditions, which are the same as those that apply for the validity of the Cramer-Rao inequality described below).

Example. $N(\mu, \sigma^2)$:

$$p(x|\,\mu, \sigma) = \exp\left\{\mu\sigma^{-2}x - \frac{1}{2}\sigma^{-2}x^2 - \left(\frac{1}{2}\mu^2\sigma^{-2} + \ln\sigma + \frac{1}{2}\ln(2\pi)\right)\right\},$$

and $B_1(\mu, \sigma) = \mu\sigma^{-2}$, $B_2(\mu, \sigma) = -\frac{1}{2}\sigma^{-2}$, $A_1(x) = x$, $A_2(x) = x^2$. The vector $S = (\sum_i x_i, \sum_i x_i^2)$ based on sample $(x_1, \ldots, x_n)$ is sufficient for $\vec{\theta} = (\mu, \sigma)$.

2.7 Estimators

Recall that an estimator is a statistic (i.e., a function of data only) that is used to obtain an estimate of one or more parameters of the underlying distribution. Often we consider *point estimators* which are single valued functions $\widehat{\theta}(X_1, \ldots, X_n)$ of $X_1, \ldots, X_n$.

Examples of point estimators:

1. if $\theta = \mathbb{E}(X)$, we can take $\widehat{\theta}$ to be mean, median, mode of the empirical distribution;

2. moment estimators, including the **sample mean**

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

   and the **sample variance**

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \, .$$

3. MLE - maximum likelihood estimator, which minimizes the *score*.

Typically there will be several possible estimators of a parameter θ. To choose between estimators we will define various desirable properties: *unbiasedness*, *consistency* and *efficiency*. *Admissibility* and *sufficiency* are also desirable properties but we won't discuss these here. Sufficiency of an estimator is closely related to sufficiency of a statistic. Robustness and ease of computation are not considered in this course, but may be important in practical applications.

2.7.1 Unbiasedness

Definition 2.1. *$\widehat{\theta}$ (r.v.) is an unbiased estimator of θ iff*

$$\mathbb{E}(\widehat{\theta}) = \theta.$$

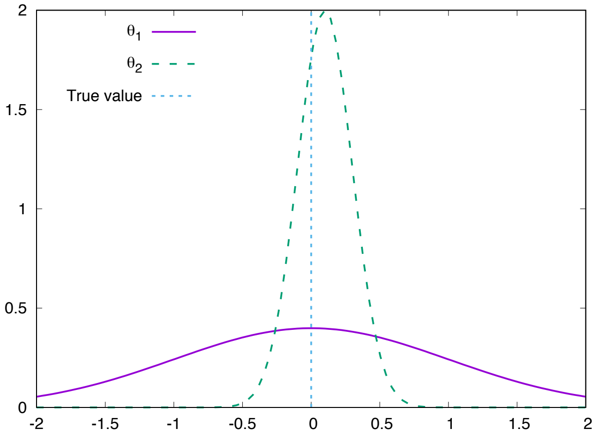
If $\mathbb{E}(\widehat{\theta}) \neq \theta$ then $\widehat{\theta}$ is a biased estimator and we define the bias function of $\widehat{\theta}$ as

$$\text{bias}(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta.$$

As an example, suppose θ is a population mean, then the sample mean $\bar{X}$ is unbiased. Also, X_1 (first observation in sample) is unbiased, and if the distribution is symmetric so is the sample median.

There are often several unbiased estimators to choose from, but which is best?

Unbiasedness is not necessarily required for all estimation problems, e.g.,



$\widehat{\theta}_1$ (with wide density) and
$\widehat{\theta}_2$ (with narrow density)
are estimators of θ;
$\widehat{\theta}_1$ is unbiased;
$\widehat{\theta}_2$ is biased;
but $\widehat{\theta}_2$ may be preferred because it is less likely to be a long way from θ.

Biased estimators may be preferred to unbiased estimators in some circumstances. A good property is asymptotic unbiasedness.

Definition 2.2. $\widehat{\theta}$ *(r.v.) is asymptotically unbiased estimator of* θ *iff*

$$\mathbb{E}(\widehat{\theta}) \to \theta \quad as \quad n \to \infty.$$

2.7.2 Consistency

As sample size is increased the sampling pdf of any reasonable estimator should become more closely concentrated about θ.

Definition 2.3. $\widehat{\theta}$ *is a (weakly) consistent estimator for* θ *if*

$$\mathbb{P}(\mid \widehat{\theta} - \theta \mid > \epsilon) \to 0 \qquad \text{as } n \to \infty$$

for any $\epsilon > 0$.

For a particular problem, it may be difficult to verify consistency from this definition, however, a sufficient (not necessary) condition for consistency is given in the lemma below.

Lemma 2.2. *If* var $(\widehat{\theta}) \to 0$ *and* bias$(\widehat{\theta}) \to 0$ *as* $n \to \infty$, *then* $\widehat{\theta}$ *is (weakly) consistent.*

Definition 2.4. *The mean square error of an estimator* $\widehat{\theta}$ *is defined as*

$$\text{MSE}(\widehat{\theta}) \quad = \quad \mathbb{E}[(\widehat{\theta} - \theta)^2] = \text{var}(\widehat{\theta}) + [\text{bias}(\widehat{\theta})]^2.$$

Mean squared error consists of two terms: variance of $\widehat{\theta}$ and its squared bias.

The *Markov inequality* states that, for a non-negative random variable X and $a > 0$

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

which can be proved straightforwardly

$$\mathbb{E}(X) = \int_0^\infty xp(x)\mathrm{d}x = \int_0^a xp(x)\mathrm{d}x + \int_a^\infty xp(x)\mathrm{d}x \geq \int_a^\infty xp(x)\mathrm{d}x \geq a\int_a^\infty p(x)\mathrm{d}x = a\mathbb{P}(X \geq a).$$

Setting $X = (\hat{\theta} - \theta)^2$ and $a = \epsilon^2$ we find

$$\mathbb{P}[\mid \widehat{\theta} - \theta \mid > \epsilon] \quad \leq \quad \frac{1}{\epsilon^2}\mathbb{E}(\widehat{\theta} - \theta)^2.$$

The term on the right had side is the mean square error. If both bias and variance tend to zero asymptotically, the mean square error tends to zero and therefore the left hand side must tend to zero. Hence we have proven Lemma 2.2.

Examples

1. Estimation of the mean of a normal distribution: using the sample mean $\bar{X}$ or median or just the value of X_1 (first observation in sample) are all unbiased estimators and have variances $\frac{\sigma^2}{n}$, $\alpha\frac{\sigma^2}{n}$ (α is a constant > 1) and σ^2. Therefore the first two are consistent. However, it is evident that X_1 is not consistent as its distribution does not change with sample size.

2. The Cauchy distribution with scale 1 and pdf $p(x|\,\theta) = \pi^{-1}[1+(x-x_0)^2]^{-1}$. In this case, the sample mean $\bar{X}$ has the same distribution as any single X_i, thus $\mathbb{P}[\mid \bar{X} - x_0 \mid > \epsilon]$ is the same for any n. This does not tend to zero as $n \to \infty$, and so $\bar{X}$ is not (weakly) consistent. (However, the sample median is a consistent estimator of x_0.)

2.8 Efficiency

Definition 2.5. *The* ***efficiency*** *of an unbiased estimator* $(\widehat{\theta})$ *is the ratio of the minimum possible variance to var*$(\widehat{\theta})$.

Definition 2.6. *An unbiased estimator with efficiency equal to 1 is called* ***efficient*** *or a* ***minimum variance unbiased estimator (MVUE)****.*

We can also define asymptotic efficiency of an (asymptotically) unbiased estimator $(\widehat{\theta})$ is the limit of the ratio of the minimum possible variance to var$(\widehat{\theta})$ as sample size $n \to \infty$.

Definition 2.7. *An estimator with asymptotic efficiency equal to 1 is called* ***asymptotically efficient****.*

We can compare the efficiency of two estimators in the following way.

Definition 2.8. *The* ***(asymptotic) relative efficiency*** *of two unbiased estimators* $\widehat{\theta}_1$ *and* $\widehat{\theta}_2$ *is the reciprocal of the ratio of their variances, as sample size* $\to \infty$*:* $\lim_{n\to\infty} \frac{Var(\widehat{\theta}_1)}{Var(\widehat{\theta}_2)}$.

The definition of asymptotic relative efficiency can also be extended to asymptotically unbiased estimators. These definitions are all fine, but they rely on knowing what the smallest possible variance is. Under certain assumptions we can obtain this from the Cramér-Rao inequality.

2.8.1 Cramér-Rao lower bound (inequality)

The theorem below (Cramér-Rao inequality) provides a lower bound on the variance of an estimator. When this lower bound is attainable for unbiased estimators, it can be used in the definition of efficiency.

Regularity conditions for the Cramér-Rao inequality.

1. $\forall\theta_1,\, \theta_2 \in \Theta$ such that $\theta_1 \neq \theta_2$, $p(x \mid \theta_1) \neq p(x \mid \theta_2)$ [identifiability].
2. $\forall\theta \in \Theta$, $p(x \mid \theta)$ have common support.
3. Θ is an open set.
4. $\exists \partial p(x \mid \theta)/\partial\theta$.
5. $\mathbb{E}\left(\partial \log p(\mathbf{X}|\theta)/\partial\theta\right)^2 < \infty$.

Here $I(\theta) = \mathbb{E}\left(\frac{\partial \log f(\mathbf{X}|\theta)}{\partial\theta}\right)^2$ is the Fisher information matrix.

Theorem 2.2. *(Cramér-Rao inequality) Let $X_1, \ldots, X_n$ denote a random sample from $p(x|\,\theta)$, and suppose that $\widehat{\theta}$ is an estimator for θ. Then, subject to the above regularity conditions,*

$$\mathrm{var}(\widehat{\theta}) \geq \frac{\left(1 + \frac{\partial b}{\partial\theta}\right)^2}{I_\theta},$$

where

$$b(\theta) = \mathrm{bias}(\widehat{\theta}) \quad \text{and} \quad I_\theta = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial\theta}\right)^2\right].$$

Comments

1. For unbiased $\widehat{\theta}$, the lower bound simplifies to $\mathrm{var}(\widehat{\theta}) \geq I_\theta^{-1}$.
2. I_θ is called Fisher's information about θ contained in the observations.
3. Regularity conditions are needed to change the order of differentiation and integration in the proof given below.
4. The result can be extended to estimators of functions of θ.

Proof of Theorem 2.2.

$$\begin{aligned}\mathbb{E}[\widehat{\theta}] &= \int \ldots \int \widehat{\theta}(x_1, \ldots, x_n) \left\{\prod_{i=1}^{n} p(x_i|\,\theta)\right\} d\mathbf{x} \\ &= \int \ldots \int \widehat{\theta}(x_1, x_2, \ldots, x_n) L(\theta; \mathbf{x}) d\mathbf{x}\end{aligned}$$

$\int \ldots \int$ is a multiple integral with respect to $\mathbf{x}$=$(x_1, x_2, \ldots, x_n)$.

From the definition of bias we have

$$\theta + b = \mathbb{E}(\widehat{\theta}) = \int \ldots \int \widehat{\theta} L(\theta; \mathbf{x}) d\mathbf{x}.$$

Differentiating both sides with respect to θ gives (using regularity conditions)

$$1+\frac{\partial b}{\partial \theta}=\int\dots\int\widehat{\theta}\frac{\partial L}{\partial \theta}d\mathbf{x}$$

since $\widehat{\theta}$ does not depend on θ. Since $l=\ln(L)$ we have

$$\frac{\partial l}{\partial \theta}=\frac{\partial \ln(L)}{\partial \theta}=\frac{1}{L}\frac{\partial L}{\partial \theta}, \quad \text{and thus} \quad \frac{\partial L}{\partial \theta}=L\frac{\partial l}{\partial \theta}.$$

Thus

$$1+\frac{\partial b}{\partial \theta}=\int\dots\int\widehat{\theta}\frac{\partial l}{\partial \theta}Ld\mathbf{x}=\mathbb{E}\left(\widehat{\theta}\frac{\partial l}{\partial \theta}\right).$$

Now use the result that for any two r.v.s $U$and V,

$$\{\mathrm{cov}(U,V)\}^2\le \mathrm{var}(U)\mathrm{var}(V)$$

and let

$$U=\widehat{\theta}, \text{ and } V=\partial l/\partial\theta.$$

Then

$$\begin{aligned}\mathbb{E}[V] &= \int\dots\int\frac{\partial l}{\partial \theta}Ld\mathbf{x}=\int\dots\int\frac{\partial L}{\partial \theta}d\mathbf{x}\\ &= \frac{\partial}{\partial\theta}\left(\int\dots\int L\ d\mathbf{x}\right) \quad \text{(using regularity conditions)}\\ &= \frac{\partial}{\partial\theta}(1)=0.\end{aligned}$$

Hence

$$\mathrm{cov}(U,V)=\mathbb{E}(UV)=1+\frac{\partial b}{\partial \theta}.$$

Similarly

$$\mathrm{var}(V)=\mathbb{E}(V^2)=\mathbb{E}\left[\left(\frac{\partial l}{\partial\theta}\right)^2\right]=I_\theta \quad \text{(by definition of } I_\theta)$$

and since $\mathrm{var}(U)=\mathrm{var}(\widehat{\theta})$ we obtain the Cramér-Rao lower bound as

$$\mathrm{var}(\widehat{\theta})\ge\frac{\{\mathrm{cov}(U,V)\}^2}{\mathrm{var}(V)}=\frac{\left(1+\frac{\partial b}{\partial\theta}\right)^2}{I_\theta}.$$

□

The Cramér-Rao lower bound will only be useful if it is attainable or at least nearly attainable.

Lemma 2.3. *The Cramér-Rao lower bound is attainable iff there exists a function $f(x)$ of x only, and functions $a(\theta)$, $c(\theta)$ of θ only such that*

$$\frac{\partial l}{\partial \theta} = \frac{(f(x) - a(\theta))}{c(\theta)},$$

in which case $\widehat{\theta} = f(x)$ attains it. The expectation value $\mathbb{E}_\theta \hat{\theta} = a(\theta)$ and $\mathrm{d}a/\mathrm{d}\theta = c(\theta) I_\theta$.

Corollary 2.1. *There is an unbiased estimator that attains the Cramér-Rao lower bound iff there exists a function $g(x)$ of x only such that*

$$\frac{\partial l}{\partial \theta} = I_\theta (g(x) - \theta),$$

in which case the unbiased estimator $\widehat{\theta} = g(x)$ attains it.

Lemma 2.4. *Under the same regularity conditions as for the Cramér-Rao lower bound*

$$I_\theta = -\mathbb{E}\left[\frac{\partial^2 l}{\partial \theta^2}\right]$$

Example
$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$, σ^2 known.
Likelihood for μ

$$L(\mu; \mathbf{x}) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\}$$

log likelihood for μ

$$l = \log L = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Thus we have

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu), \qquad \frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2},$$

and

$$I_\theta = \mathbb{E}\left[-\frac{\partial^2 l}{\partial \mu^2}\right] = \frac{n}{\sigma^2}.$$

The lower bound for unbiased estimators is $I_\theta^{-1} = \frac{\sigma^2}{n}$. However,

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n},$$

so $\bar{X}$ attains its lower bound. No other unbiased estimator can have smaller variance than $\bar{X}$. Therefore $\bar{X}$ is MVUE.

Alternatively, we can use Lemma 2.3, and

$$\frac{\partial l}{\partial \mu} \quad = \quad \frac{1}{\sigma^2}\sum (X_i - \mu) = \frac{n}{\sigma^2}(\bar{X} - \mu)$$

Therefore the bound is attainable.

Regularity conditions are essential to be able to use the lower bound. Consider the uniform distribution case $X_1, X_2, \ldots, X_n \sim U[0, \theta]$

$$L(\theta; \mathbf{x}) = \begin{cases} \frac{1}{\theta^n} & 0 \leq x_{(1)} \leq x_{(2)} \leq \ldots, \leq x_{(n)} \leq \theta \\ 0 & \text{elsewhere} \end{cases}$$

In the range where L is differentiable $l = -n \log \theta$

$$\frac{\partial l}{\partial \theta} = -\frac{n}{\theta} \quad \text{and} \quad \frac{\partial^2 l}{\partial \theta^2} = \frac{n}{\theta^2}.$$

Thus

$$I_\theta = \mathbb{E}\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] = \frac{n^2}{\theta^2}$$

but

$$\mathbb{E}\left[-\frac{\partial^2 l}{\partial \theta^2}\right] = \frac{-n}{\theta^2}.$$

Therefore the lower bound should be $\frac{\theta^2}{n^2}$, but

$$\text{var}\left[\frac{n+1}{n} X_{(n)}\right] = \frac{\theta^2}{n(n+2)} < I_\theta^{-1}.$$

The lower bound is violated because the regularity conditions don't hold. In particular the second condition is violated, since the support of the distribution depends on θ.

The derivation and examples above were all for a one dimensional parameter. The corresponding result for the multiple parameter case is

$$\text{cov}(t_i, t_j) \geq \frac{\partial m_i}{\partial \theta_k} \left[\mathbf{I}_\theta\right]^{-1}_{kl} \frac{\partial m_j}{\partial \theta_l}, \qquad \left[\mathbf{I}_\theta\right]_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j}\right],$$

where $\mathbf{t}$ is the realised value of some multi-dimensional statistic $\mathbf{T}$ and $\mathbf{m} = \vec{\theta} + \mathbf{b} = \mathbb{E}(\mathbf{T})$.

2.9 Rao-Blackwell Theorem

The Rao-Blackwell theorem gives a method of improving an unbiased estimator, and involves conditioning on a sufficient statistic.

Theorem 2.3. *(Rao-Blackwell theorem). Let X_1, $X_2, \ldots, X_n$ be a random sample of observations from a distribution with pdf $p(x|\,\theta)$. Suppose that S is a sufficient statistic for θ and that $\widehat{\theta}$ is any unbiased estimator for θ. Define $\widehat{\theta}_S = \mathbb{E}[\widehat{\theta} \mid S]$. Then*

(a) $\widehat{\theta}_S$ is a function of S only;

(b) $\mathbb{E}[\widehat{\theta}_S] = \theta$;

(c) $\text{var}\, \widehat{\theta}_S \leq \text{var}\, \widehat{\theta}$.

2.10 Maximum likelihood estimators

Definition 2.9. *The maximum likelihood estimator (MLE) is defined by* $\widehat{\theta} = \arg\max_{\theta\in\Theta} L(\theta;\mathbf{x}) = \arg\max_{\theta\in\Theta} \ell(\theta;\mathbf{x})$.

If $\exists \partial\ell/\partial\theta_j$ and Θ is open, then the MLE $\widehat{\theta}$ satisfies $\partial\ell/\partial\theta_j(\widehat{\theta}) = 0$, $j = 1,\ldots,K$, $\theta \in \Theta \subset \mathbb{R}^K$.

The MLE can be biased or unbiased but it is asymptotically unbiased and efficient and it is also consistent. In fact the following lemma holds.

Lemma 2.5. *Let* $X_1,\ldots,X_n \sim p(x \mid \theta)$ *IID,* $\theta \in \Theta \subset \mathbb{R}^K$*. Under the regularity conditions of Cramer-Rao inequality, the MLE asymptotically satisfies*

$$\widehat{\theta} \sim N_K(\theta, I_\theta^{-1}) \quad n \to \infty,$$

in particular, $\mathbb{E}(\widehat{\theta}) \to \theta$ *and for* $K = 1$, $Var(\widehat{\theta})/I_\theta^{-1} \to 1$ *as* $n \to \infty$.

If there exists an unbiased efficient estimator this has to be the MLE.

Lemma 2.6. *Suppose there exists an unbiased estimator* $\tilde{\theta}$ *that attains Cramer-Rao lower bound, and suppose that MLE* $\hat{\theta}$ *is the solution of* $\frac{\partial\ell}{\partial\theta} = 0$*. Then,* $\tilde{\theta} = \hat{\theta}$*.*

Proof. $\tilde{\theta}$ is unbiased and attains Cramer-Rao lower bound, hence, by the corollary to Lemma 2.3, $\frac{\partial\ell}{\partial\theta} = I_\theta(\tilde{\theta} - \theta)$. Then, the only solution of $\frac{\partial\ell}{\partial\theta} = 0$ is $\tilde{\theta}$, that is, $\tilde{\theta} = \hat{\theta}$. □

Thus, (under the regularity conditions of Cramer-Rao inequality) if the Cramer-Rao lower bound is attainable, the MLE attains it, thus in this case the MLE is efficient. If the bound is unattainable, then the MLE is asymptotically efficient.

2.11 Confidence intervals and regions

Point estimators provide single estimated values for parameters, but we usually also need an estimate of the uncertainty in those estimated values. These are characterised by **confidence intervals**. A confidence interval is a random variable since the ends of the interval are typically determined as a function of the observed data. The interval has the property that over many realisations of the same experiment, the intervals constructed randomly by this procedure will contain the true value of the parameter a certain fraction of the time.

Formally a set $S_\alpha(\mathbf{X})$ is a $(1-\alpha)$ **confidence region** for ψ if

$$\mathbb{P}(S_\alpha(\mathbf{X}) \ni \psi; \psi, \lambda) = 1 - \alpha \quad \forall \psi, \lambda.$$

Thus, $S_\alpha(\mathbf{X})$ is a random set of ψ-values which includes the true value with probability $1-\alpha$. If more than one value of α is considered, we usually require

$$S_{\alpha_1}(\mathbf{x}) \supset S_{\alpha_2}(\mathbf{x}) \text{ if } \alpha_1 < \alpha_2. \tag{53}$$

e.g. a 99% region contains the 95% region.

If ψ is a scalar and $S_\alpha(\mathbf{x})$ has the form $\{\psi : t^\alpha \geq \psi\}$ for some statistic t^α, then t^α is a $(1-\alpha)$ **upper confidence limit** for ψ.

If ψ is a scalar and $S_\alpha(\mathbf{x})$ has the form $\{\psi : s^\alpha \leq \psi\}$ for some statistic s^α, then s^α is a α **lower confidence limit** for ψ.

If $S_\alpha(\mathbf{x}) = \{\psi : a_\alpha(\mathbf{x}) \leq \psi \leq b_\alpha(\mathbf{x})\}$, it is a **two-sided confidence interval**.

A two-sided confidence interval is called **equitailed** if $a_\alpha(\mathbf{x})$ is the $\alpha/2$ lower confidence limit and $b_\alpha(\mathbf{x})$ is the $1-\alpha/2$ upper confidence limit.

A **high density confidence region** is $\{\theta \in \Theta : p(\mathbf{x}|\theta) \geq K_\alpha\}$ where the constant K_α is determined by the condition $\mathbb{P}\{p(\mathbf{X}|\theta) \geq K_\alpha\} = 1-\alpha$.

Confidence intervals/regions for estimators can be constructed by identifying **pivotal quantities**. A pivotal quantity $U = u(\mathbf{X}, \psi)$ is a scalar function of $\mathbf{X}$ and ψ with the same distribution for all ψ and λ. If u_α is the upper α point of this distribution, then

$$\mathbb{P}(u(\mathbf{X}, \psi) \leq u_\alpha) = 1 - \alpha,$$

so that the set $\{\psi : u(\mathbf{x}, \psi) \leq u_\alpha\}$ defines a $(1-\alpha)$ confidence region for ψ.

If ψ is a scalar and $u(mathbfx, \psi)$ is monotone in ψ, this yields a one-sided interval. In this case we may also define two-sided intervals by $\{\psi : u_{\alpha_L} \leq u(\mathbf{x}, \psi) \leq u_{\alpha_U}\}$ with $\alpha_U - \alpha_L = 1 - \alpha$.

Examples of pivotal quantities

- $\mathcal{E}(\lambda)$: $2\theta \sum X_j$ which has distribution $\chi^2(2n)$;
- $N(\mu, \sigma^2)$, inference about μ with σ unknown: $\sqrt{n}(\bar{x} - \mu)/s$ which has distribution $t(n-1)$;
- Ratio of two Normal variances: $(s_1^2/\sigma_1^2)/(s_2^2/\sigma_2^2)$ which has distribution $F(n_1-1, n_2-1)$.

3 Hypothesis testing

Often when we observed data we have some ideas about the random processes that are generating the observations. Having collected data it is natural to test whether the observed data are consistent with those expectations. The idea of hypothesis testing is to say if the data provides sufficient evidence to rule out those assumptions. The emphasis is always placed in favour of the assumptions, rather than the alternative. We require strong evidence that the data are inconsistent with the assumptions before we reject them.

Formally, we suppose that we have data $\mathbf{x} = (x_1, \ldots, x_n)$ and want to examine whether they are consistent with a hypothesis H_0 (the **null hypothesis** or **hypothesis under test**) about the distribution function $F_{\mathbf{X}}$ of $\mathbf{X}$.

A hypothesis is **simple** if it defines $P_{\mathbf{X}}$ completely:

$$H_0 : P_{\mathbf{X}} = P_0$$

otherwise, it is **composite**. If $P_{\mathbf{X}}$ is parametric with more than one parameter, a composite hypothesis might specify the values of some or all of them. (e.g. one regression coefficient)

The distribution of $\mathbf{X}$ under H_0, P_0, is called null distribution.

Examples of hypotheses

- A significant trigger in a gravitational wave detector is due to instrumental fluctuations. This is a composite hypothesis as the distribution of triggers under the noise assumption is not fully specified.

- The numbers of gravitational wave events $x_1, \ldots, x_7$ observed on Monday, ..., Sunday. The null hypothesis is that all days are equally likely, i.e., the joint distribution is Multinomial$(n; \frac{1}{7}, \ldots, \frac{1}{7})$. This is a simple hypothesis.

- The right ascensions $x_1, \ldots, x_n$ angles of observed gravitational wave events. The hyypothesis that the X_j's are independently Uniform on $[0, 2\pi)$ is simple.

  Suppose we want to test that there is clustering around some angle, then we can assume that the distribution is von Mises with pdf

  $$p(x|\,\theta, \lambda) = \frac{1}{2\pi I_0(\lambda)} e^{\lambda \cos(x-\theta)}, \quad x \in \mathcal{X} = [0, 2\pi);\ \lambda \geq 0,\ 0 \leq \theta < 2\pi;$$

  for unknown λ. This is a composite hypothesis.

- The hypothesis that the number of gravitational wave events in each month $X_1, \ldots, X_n$ are independently Poisson(θ) with unknown θ is composite.

3.1 Definitions and basic concepts

1. A sample of n observations is available to make inference about parameter θ.

2. We wish to decide between two hypotheses: H_0, *the null hypothesis*, and H_1, *the alternative hypothesis.*

   H_0 is often *simple* (only one value is specified for θ)

   $$\text{i.e. } H_0 : \theta = \theta_0 \text{ (e.g. } H_0 : \mu = 100,\ H_0 : p = \tfrac{1}{2}\text{)}.$$

H_1 can be simple: $H_1 : \theta = \theta_1$ but more commonly it is *composite* (more than one value is allowed for θ). The most common alternatives are

$$H_1 : \theta < \theta_0 \;\text{ or }\; H_1 : \theta > \theta_0 \text{ — \underline{one-sided/one-tailed alternative}}$$
$$\text{or } H_1 : \theta \neq \theta_0 \text{ — \underline{two-sided/two-tailed alternative}.}$$

3. Two possible decisions: *to reject* or *not to reject* H_0 in favour of H_1.

   The decision whether or not to reject H_0 is based on the value of a *test statistic*, which is a function of the observations.

4. Values of the test statistic for which H_0 is not rejected form the *acceptance region*, $\bar{C}$.

   Values of the test statistic for which H_0 is rejected form the *rejection region* (or critical region), C.

   The form of these regions depends on the form of H_1.

5. There are two possible types of error:

   Reject H_0 when H_0 is true — Type I error
   Fail to reject H_0 when H_0 is false — Type II error

   The probability of Type I error, denoted by α, is the **significance level (**or **size)** of the test.

   The probability of Type II error, denoted by β, is only defined uniquely if H_1 is simple. In which case

   $$\eta \;=\; 1 - \beta \text{ is the \textbf{power} of the test.}$$

   For composite H_1, $\eta(\theta)$ is the *power function*.

Generally we consider Type-I error (false rejection) to be worse than Type-II (incorrect failure to reject) as usually in the latter case more data will be collected and the test will be re-evaluated. It is therefore usual to specify the **significance level** of the test in order to determine the threshold for rejection, or the quote a **p-value** (see next section) when quoting test results.

We can define a **test function** $\phi(x)$ such that

$$\phi(x) = \begin{cases} 1 & \text{if } t(\mathbf{x}) \in C \\ 0 & \text{if } t(\mathbf{x}) \in \bar{C} \end{cases}$$

and when we observe $\phi(\mathbf{X}) = 1$, we reject H_0. This function has the property that $\alpha = \mathbb{E}_{H_0}(\phi(\mathbf{X}))$ and $\eta = \mathbb{E}_{H_1}(\phi(\mathbf{X}))$, in which the subscript denotes the hypothesis under which the expectation value is to be calculated.

For discrete distributions, the probability that the test statistic lies on the boundary of the critical region, ∂C, may be non-zero. In that case, it is sometimes necessary to use a **randomized test**, for which the test function is

$$\phi(x) = \begin{cases} 1 & \text{if } t(\mathbf{x}) \in C \\ \gamma(\mathbf{x}) & \text{if } t(\mathbf{x}) \in \partial C \\ 0 & \text{if } t(\mathbf{x}) \in \bar{C} \end{cases}$$

for some function $\gamma(\mathbf{x})$ and we reject H_0 based on observed data $\mathbf{x}$ with probability $\phi(\mathbf{x})$.

3.2 Test statistic

Often to construct a test (i.e. the decision whether to reject H_0 or not based on observed data $\mathbf{x}$), a test statistic is used.

Definition 3.1. *A real-valued function $t(\mathbf{x})$ on $\mathcal{X}$ is a test statistic for testing H_0 iff*

(i) values of t are **ordered** *with respect to the evidence for departure from H_0*

(ii) the distribution of $T = t(\mathbf{X})$ under H_0 is known, at least approximately. For composite H_0 the distribution should be (approximately) the same for all simple hypotheses making up H_0.

For any observation $\mathbf{x}$, we measure the consistency of $\mathbf{x}$ with H_0 using the *significance probability* or the *p-value*, e.g. if larger values of t correspond to stronger evidence for departure from H_0, the p-value is defined by

$$p = \mathbb{P}(T \geq t(\mathbf{x}) |\, H_0),$$

the probability (under H_0) of seeing the observed value of t or any more extreme value. The smaller the value of p the greater the evidence against H_0.

3.3 Alternative hypothesis

Can be specified or unspecified.

3.3.1 Pure significance tests

In a *pure significance test*, only the null hypothesis H_0 is explicitly specified. The p-value of the observed value under the null distribution is evaluated, and if it is sufficiently small, the null hypothesis would be rejected. Such tests are done if we want to avoid specifying a parametric family of alternative distributions.

There will often be multiple quantities that could be computed under the null hypothesis and we can choose any of them to evaluated the distribution of the test statistic. The best choice can be guided if we have a specific idea of the type of departure from H_0 we are looking for, e.g.,

- Directional data: Might look for a tendency for the observed directions to cluster about a (possibly unknown) direction. But not a specific set of alternatives such as von Mises distributions.

- Pois(θ): if the alternative is not a Poisson distribution, we might test whether variance $\neq$ expectation.

An important class of pure significance tests are *goodness of fit* tests where either the sample distribution function $\hat{P}_X(x) = \frac{1}{n}\sum_{i=1}^{n} I(x \leqslant x_i)$ or the histogram are compared to those of the null distribution.

Examples

- Event frequency on different days: $H_0: \ X_1, \ldots, X_7 \sim \text{Mult}(n; \frac{1}{7}, \ldots, \frac{1}{7})$.

  With no particular alternative we might use Pearson's χ^2 test, comparing

  $$X^2 = \sum_{i=1}^{7} \frac{\left(x_i - \frac{n}{7}\right)^2}{\frac{n}{7}} \quad \text{with} \quad \chi^2_6.$$

- Right ascension of GW sources: If alternative to H_0 is clustering about the reference direction (e.g. galactic centre) we could use $\sum \cos x_j$, the projection onto the reference axis of the resultant sum vector $(\sum \cos x_j, \sum \sin x_j)$.

- $\text{Pois}(\theta)$: might use index of dispersion,

  $$d = \frac{\sum (x_i - \bar{y})^2}{\bar{y}},$$

  which is approximately χ^2 with $(n-1)$ degrees of freedom under H_0 for $\theta \geq 1$.

  Note that given $\sum X_j = s$, the distribution of $X_1, \ \ldots \ , X_n$ is $\text{Mult}(s, \frac{1}{n}, \ \ldots \ , \frac{1}{n})$ and d is the χ^2 statistic for testing the fit of this distribution.

3.3.2 Specified alternative hypothesis

For a parametrised family of distributions $p(x|\,\theta)$, $\theta \in \Theta$, say $H_0 : \ \theta = \theta_0$, then

$$H_1 : \ \theta \in \Theta_1 \subset \Theta \setminus \{\theta_0\},$$

e.g. $\theta \neq \theta_0$ (two-sided), $\theta > \theta_0$ or $\theta < \theta_0$ (one-sided).

Below we consider two cases: with simple and composite alternative hypotheses (and a simple null hypothesis).

With composite alternative hypotheses, the power of the test becomes the power function defined over $\theta \in \Theta_1$:

$$\eta(\theta) = \mathbb{P}(\text{reject } \ H_0|\,\theta) = \mathbb{P}_\theta(\text{reject } \ H_0).$$

3.4 Critical regions

In § 3.2 we defined for each $\mathbf{x} \in X$ the significance probability

$$p = \mathbb{P}(T \geq t(\mathbf{x})|\,H_0)$$

associated with a test statistic t. A different, but equivalent, approach defines a test using critical regions rather than test statistics. This

(i) facilitates comparison of different tests of H_0 according to their properties under H_1;

(ii) is useful for establishing a connection between tests and confidence regions.

For any α in the interval $(0, 1)$, a subset R_α of X is a **critical region of size** α if

$$\mathbb{P}(\mathbf{X} \in R_\alpha|\,H_0) = \alpha \tag{54}$$

Interpretations of R_α:

(i) points in R_α are regarded as not consistent with H_0 at level α;

(ii) points in R_α are "significant at level α";

(iii) if $\mathbf{x} \in R_\alpha$, then H_0 is "rejected" in a test of size α.

A significance test is defined by a set of critical regions $\{R_\alpha : 0 < \alpha < 1\}$ satisfying

$$R_{\alpha_1} \subset R_{\alpha_2} \;\text{ if } \alpha_1 < \alpha_2. \tag{55}$$

Thus, for example, if data $\mathbf{x}$ are significant at the 1% level, they are also significant at the 5% level.

The **significance probability** (also called p-value) for data $\mathbf{x}$ is then defined as

$$P = \inf(\alpha; \mathbf{x} \in R_\alpha),$$

i.e. the smallest α for which $\mathbf{x}$ is significant at level α.

The definition of a test in §3.2 corresponds to critical regions of the form

$$R^t_\alpha = \{\mathbf{x} : t(\mathbf{x}) \geq t_\alpha\},$$

where t_α is the upper α point of $T = t(\mathbf{X})$ under H_0, since

$$\mathbb{P}(\mathbf{X} \in R^t_\alpha | \, H_0) = \mathbb{P}(t(X) \geq t_\alpha | \, H_0) = \alpha,$$

by the definition of t_α; also if $\alpha_1 < \alpha_2$ then $t_{\alpha_1} > t_{\alpha_2}$ and $R^t_{\alpha_1} \subset R^t_{\alpha_2}$ satisfying (55). Finally,

$$\begin{aligned} P &= \mathbb{P}(t(\mathbf{X}) \geq t(\mathbf{x}) : H_0) \\ &= \inf(\alpha; t(\mathbf{x}) \geq t_\alpha) \\ &= \inf(\alpha; \mathbf{x} \in R^t_\alpha), \end{aligned}$$

the smallest α for which $\mathbf{x}$ is significant at level α.

Example

- X_j independent $N(\mu, \sigma^2)$ (σ known and hence =1 without loss of generality) To test $H_0 : \mu = \mu_0$ vs $\mu > \mu_0$, obvious test statistics are $\bar{Y}$ or $(\bar{Y} - \mu_0)\sqrt{n}$. The significance probability is

$$P = \mathbb{P}\left((\bar{Y} - \mu_0)\sqrt{n} > (\bar{y} - \mu_0)\sqrt{n} | \, H_0\right) = 1 - \Phi((\bar{y} - \mu_0)\sqrt{n}).$$

The corresponding critical regions are $R_\alpha = \{\mathbf{x} : (\bar{y} - \mu_0)\sqrt{n} \geq \Phi^{-1}(1-\alpha)\}$. Thus

$$\mathbb{P}(\mathbf{X} \in R_\alpha | \, H_0) = \mathbb{P}((\bar{Y} - \mu_0)\sqrt{n} \geq \Phi^{-1}(1-\alpha)) = \alpha,$$

as required, and if $\alpha_1 < \alpha_2$, then $\Phi^{-1}(1-\alpha_1) > \Phi^{-1}(1-\alpha_2)$, so that $R_{\alpha_1} \subset R_{\alpha_2}$. Also

$$\begin{aligned} \inf(\alpha; \mathbf{x} \in R_\alpha) &= \inf(\alpha; (\bar{y} - \mu_0)\sqrt{n} \geq \Phi^{-1}(1-\alpha)) \\ &= \inf(\alpha; \alpha \geq 1 - \Phi((\bar{y} - \mu_0)\sqrt{n}) \\ &= P. \end{aligned}$$

3.5 Construction of confidence intervals using critical regions

The construction of hypothesis tests leads naturally to the construction of confidence intervals and regions. For any value ψ_0 of ψ, let $R_\alpha(\psi_0)$ be a size-α critical region for testing the null hypothesis $\psi = \psi_0$ against $\psi \neq \psi_0$ (or possibly $\psi < \psi_0$ or $\psi > \psi_0$). For any $\mathbf{x}$ define

$$S_\alpha(\mathbf{x}) = \{\psi_0 : \mathbf{x} \notin R_\alpha(\psi_0)\}.$$

Then $S_\alpha(\mathbf{X})$ is a $(1-\alpha)$ confidence interval for ψ since

$$\mathbb{P}(S_\alpha(\mathbf{X}) \ni \psi_0; \psi_0, \lambda) = \mathbb{P}(\mathbf{X} \notin R_\alpha(\psi_0) : \psi_0, \lambda) = 1 - \alpha \quad \forall \psi_0, \lambda$$

[$\bar{R}_\alpha(\psi_0)$ comprises $\mathbf{x}$ values judged consistent with ψ_0 (at level α), so $S_\alpha(\mathbf{x})$ comprises ψ values consistent with $\mathbf{x}$.]

If $\alpha_1 < \alpha_2$, then from (19) $\{\psi_0 : \mathbf{x} \in R_{\alpha_1}(\psi_0)\} \subset \{\psi_0 : \mathbf{x} \in R_{\alpha_2}(\psi_0)\}$, so that (53) holds.

For scalar ψ, critical regions for alternatives $\psi < \psi_0$ lead to upper confidence limits.

Example

- Exp(λ): Find the best size-α critical region for testing $\lambda = \lambda_0$ against $\lambda < \lambda_0$.

  The best size-α critical region for testing $\lambda = \lambda_0$ against $\lambda < \lambda_0$ is $R_\alpha(\lambda_0) = \{\mathbf{x} : \sum x_j > \frac{1}{2}\lambda_0^{-1}\chi^2_{2n}(\alpha)\}$. The corresponding $(1-\alpha)$ confidence region for λ is $\{\lambda_0 : \sum x_j \leq \frac{1}{2}\lambda_0^{-1}\chi^2_{2n}(\alpha)\}$ i.e. $\{\lambda_0 : \lambda_0 \leq \frac{1}{2}(\sum x_j)^{-1}\chi^2_{2n}(\alpha)\}$, so that $\frac{1}{2}(\sum x_j)^{-1}\chi^2_{2n}(\alpha)\}$ is the $(1-\alpha)$ upper confidence limit for λ.

3.6 Examples of hypothesis tests

We give three commonly encountered examples of hypothesis tests.

3.6.1 z-test

Suppose that we observe two independent samples

$$X_1, \ldots, X_n \sim N(\mu_1, \sigma^2), \qquad Y_1, \ldots, Y_m \; N(\mu_2, \sigma^2).$$

We assume additionally that σ^2 is known and we are interested in testing the hypothesis

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1 : \mu_1 - \mu_2 \neq 0.$$

If the null hypothesis is violated we expect that the magnitude of the difference in sample means, $|\bar{X} - \bar{Y}|$, will be large. The statistic

$$Z = \left(\frac{1}{n} + \frac{1}{m}\right)^{-\frac{1}{2}} \frac{(\bar{X} - \bar{Y})}{\sigma}$$

follows a $N(0,1)$ distribution under the null hypothesis so we use a critical region of the form

$$|z| > z_{\frac{\alpha}{2}}$$

to define a test with significance α. Here $z_{\frac{\alpha}{2}}$ denotes the upper $\alpha/2$ point in the Normal distribution, i.e., the point such that

$$\mathbb{P}(X \sim N(0,1) > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}.$$

3.6.2 t-test

We now suppose that we want to test the same hypothesis as in the previous example, but assuming that σ^2 is not known. Once again, we expect the difference in sample means to be large when the null hypothesis is false, but exactly how large now depends on the unknown value of σ^2. If we use the same test statistic, but with the known variance replaced by the estimated value we have

$$T = \left(\frac{1}{n} + \frac{1}{m}\right)^{-\frac{1}{2}} \frac{(\bar{X} - \bar{Y})}{\hat{\sigma}} \qquad \text{where } \hat{\sigma}^2 = \frac{1}{m+n-2}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_i - \bar{Y})^2\right)$$

which follows a t_{m+n-2} distribution under the null hypothesis.

The critical region of a size-α test is to reject H_0 when

$$|t| > t_{\frac{\alpha}{2}},$$

where $z_{\frac{\alpha}{2}}$ denotes the upper $\alpha/2$ point in the t-distribution with $m+n-2$ degrees of freedom.

3.6.3 Analysis of variance: F-test

Suppose we have observations of random variables X_{ij} where $j = 1, \ldots, n_i$ labels different observations of one particular group, and $i = 1, \ldots, k$ labels the different groups. We denote the mean in each group by

$$\bar{X}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$$

and the overall mean by

$$\bar{X}_{\bullet\bullet} = \frac{1}{N}\sum_{ij} X_{ij}, \qquad N = \sum_{i=1}^{k} n_i.$$

We are interested in testing that the means of all the groups are equal. If this is true then we expect that the **between samples sum of squares**

$$SS_b = \sum_i n_i(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2$$

is comparable to the **within samples sum of squares**

$$SS_w = \sum_{ij}(x_{ij} - x_{i\bullet})^2.$$

If the means are different then we expect the former to be larger than the latter. Therefore, we reject the null hypothesis for large values of SS_b/SS_w. The quantity

$$F = \frac{(N-k)SS_b}{(k-1)SS_w}$$

follows an $F_{k-1,N-k}$-distribution under the null hypothesis and so our critical regions are of the form to reject H_0 when

$$F > F_{k-1,N-k}(\alpha)$$

the upper α critical point of the $F_{k-1,N-k}$ distribution.

3.7 Calculating thresholds for tests

For the examples above the test statistics followed known distributions under the null hypothesis and so the critical values can be directly calculated. This is not always possible. In other situations it might be possible to compute the mean, μ, and variance, σ^2, of the test statistic, if not its full distribution. In that case, a Normal approximation can often be used by appealing to the Central Limit Theorem.

Example: $\mathcal{E}(\lambda)$: we saw above that $X = \sum x_j$ can be used for testing $\lambda = \lambda_0$ versus $\lambda < \lambda_0$. While in this case we know the exact distribution of the test statistic, if we did not we can approximate

$$X \sim N\left(\frac{n}{\lambda_0}, \frac{n}{\lambda_0^2}\right)$$

and reject the hypothesis at significance α if

$$\frac{\lambda_0 X - n}{\sqrt{n}} > z_\alpha.$$

The power of the test can be approximated in a similar way, by writing down a Normal approximation to the distribution of the test statistic under the alternative hypothesis.

If the mean and variance cannot be easily calculated, or the form of the test statistic does not lend itself to approximation by the Central Limit Theorem, then usually the best approach is to do a **simulation study**, i.e., generate many realisations of the test statistic under H_0 and determine thresholds numerically. In principle, the power of the test can be evaluated in a similar way although this might not be practical for composite alternative hypotheses.

3.8 Multiple testing

When presented with new data, there is a temptation to keep asking different questions of the same data. When doing this you have to be careful to avoid **multiple testing** (or, in the language of the gravitational wave community **trials factors**). If you keep carrying out independent tests that have a significance of α then you would expect to reject a hypothesis every $1/\alpha$ tests purely by chance. Therefore, if you plan to carry out m independent tests and want the overall significance to be α, the significance levels applied to the individual tests must be lower.

If we carry out m independent tests, each with significance α, then the combined significance is

$$1 - (1-\alpha)^m = \alpha_c.$$

To reach a target significance of the combined tests requires using individual tests with significance $\alpha = 1 - (1-\alpha_c)^{1/m} = 1 - \exp(\log(1-\alpha_c)/m) \approx \alpha_c/m$. The first expression is the *Ŝidák correction*, while the latter correction is referred to as the *Bonferroni correction.*

It is also possible to not divide the total significance evenly between the different individual tests. The *Holm-Bonferroni method* orders the individual test p-values and then tests the i'th (starting from the smallest) at a significance level of $\alpha_c/(m-i+1)$. This approach gives better overall performance.

In practice, multiple tests on the same data will not be independent and so using the corrections based on independence will be conservative and the true significance of any rejection of the null hypothesis will be greater (i.e., the true p-value will be smaller than that estimated in this way). Understanding the dependency of multiple tests is typically highly non-trivial so it is usually best to assess the true p-value of a testing programme using simulations.

Another issue to be cautious of is changing the question based on the data. Changing the question based on what was observed can lead to results appearing significant when they are not, as the following example illustrates.

Example: LIGO/Virgo operate for 8 months from January to August and sees event counts $(1, 0, 0, 0, 0, 1, 1, 4)$. Are the 4 events in the last month unusual? A total of 7 events have been observed in 8 months, so we have a rate of $\sim 7/8$ per month. Assuming that the events are Poisson distributed with this rate, the probability that a given month would have 4 or more events in it is $\sim 1.2\%$, which would be significant at the 5% level usually used for hypothesis tests. But it is not fair to ask "Is four events in August unusual?", since we only decided to look at August in particular when we saw the data. The fair question to ask is "Is four events in one of the months unusual", which means we must multiply by 8 to account for the fact that we have 8 potentially unusual months to choose from. The resulting probability of $\sim 9.8\%$ is much less significant [1]. Note that it is perfectly fine, having made these observations, to ask "Is August unusual in the next observing run?" and specifically target the month that was an outlier in previous data in the next analysis. However, this is less sensitive than doing the test "Is any month unusual?" on all of the data from both observing runs together. Suppose in the next year we also take data from January to August and observe events $(0, 1, 0, 1, 1, 0, 0, 2)$. The probability of observing two or more events in August, given the rate of 5/8 events per month, is 13%, so this would not be considered significant. However, adding the two observing runs together we have $(1, 1, 0, 1, 1, 1, 1, 6)$ and the rate for binned observations is 4/3. The probability of seeing 6 or more events in a Poisson distribution with rate 4/3 is 0.25%, which is significant [2].

3.9 Receiver operator characteristic

As mentioned above, Type-I errors are considered to be more serious than Type-II errors and so tests are quoted by the significance level. However, there may be (infinitely) many tests with the same significance, so how do we choose between them? This is done using the power function. Clearly if one test is more powerful than another for the same significance level then it is better and should be used.

In general, one way to compare different tests is by plotting a **receiver operator characteristic** (ROC) curve. This is a plot of the power versus significance of a test, or equivalently the "detection rate" of deviations in the null hypothesis against the "false alarm rate". For a random test, i.e., we toss a coin and, regardless of the observed data, say that if it is heads we have made a detection, the ROC curve is the diagonal line. Tests that lie above the line are more powerful than random at given significance, and so the further away from the diagonal line the better the test is. ROC curves can be used to compare tests visually, or

[1] Another way to tackle this problem is to say that we expect the distribution of events across the 8 months to be Multinomial with equal probability of 0.125 in each month. The distribution of events in a specific month is Binomial with $n = 7$ and $p = 0.125$ and so the probability that a specific event will have four or more events out of the 7 is $\sim 0.6\%$, but this rises to $\sim 5.0\%$ when we compute the probability that one (unspecified) month has four or more events.

[2] In the multinomial analysis the probabilities are 12% and 0.18% respectively

by computing the area between the curve and the diagonal line. Sometimnes the curves can cross, so one test may be better at one significance level and another at another. The best test then depends on what regime you are operating in.

In the following subsections we will present a number of results that describe how to find tests that have the highest power at a given significance, under various assumptions about the hypotheses and the underlying distributions. As we shall see below, it is not always possible to find a test that is the best everywhere.

3.10 Designing the best test: simple null and alternative hypotheses

Consider null and alternative hypotheses H_0, H_1 corresponding to completely specified p.d.f.'s p_0, p_1 for $\mathbf{X}$. For these hypotheses, comparison between the critical regions of different tests is in terms of

$$\mathbb{P}(\mathbf{X} \in R_\alpha | \, H_1)$$

the **power** of a size-α critical region R_α for alternative H_1. A **best** critical region of size α is one with maximum power.

In terms of p_0, p_1, the power is

$$\begin{aligned} \int_{R_\alpha} p_1(\mathbf{x}) d\mathbf{x} &= \int_{R_\alpha} p_0(\mathbf{x}) r(\mathbf{x}) d\mathbf{x} \quad \left(\text{or } \sum_{R_\alpha} p_0(\mathbf{x}) r(\mathbf{x}) \right) \\ &= \mathbb{E}\{r(\mathbf{X}) | \mathbf{X} \in R_\alpha ; H_0\} \end{aligned}$$

where

$$r(\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} = \frac{L(\theta; \mathrm{H}_1)}{L(\theta; \mathrm{H}_0)},$$

the **likelihood ratio** (LR) for H_1 vs H_0. We can **prove** that the power is maximized when R_α has the form $\{\mathbf{x} : r(\mathbf{x}) \geq k_\alpha\}$ or $\{\mathbf{x} : \frac{L(\theta;\mathrm{H}_1)}{L(\theta;\mathrm{H}_0)} \geq k_\alpha\}$, i.e. when R_α is a LR critical region. Thus we have the Neyman-Pearson lemma.

Theorem 3.1. *(Neyman-Pearson lemma). For any size α, the LR critical region is the best critical region for testing simple hypotheses H_0 vs H_1. (It is also better than any critical region of size $< \alpha$.)*

A LR test is a test whose critical regions are LR critical regions for all α for which such a size-α region exists (all α in the continuous case).

Examples

- Angles: If H_0, H_1 correspond to a Uniform distribution and a von Mises distribution with parameter θ_1, the LR is

$$r(\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} = \{2\pi I_0(\theta_1)\}^{-n} \frac{e^{\theta_1 \sum_j \cos x_j}}{(2\pi)^{-n}},$$

which is an increasing function of $t(\mathbf{x}) = \sum \cos x_j$. So the LR critical regions have the form $\{\mathbf{x} : \sum \cos x_j > t_\alpha\}$. For any α, t_α is given by $\mathbb{P}(\sum \cos X_j \geq t_\alpha | \, H_0) = \alpha$. From §3.3 $\sum \cos X_j$ is approximately $N(0, \frac{1}{2}n)$ under H_0, so t_α is approximately $\left(\frac{1}{2}n\right)^{1/2} \Phi^{-1}(1-\alpha)$. Note that the critical regions, and hence the test, do not depend on the value of θ_1.

- $\mathcal{E}(\lambda)$: $X_1, \ldots, X_n$ are i.i.d. with d.f. $1-e^{-\lambda y}$ $(y > 0)$. H_0 is $\lambda = \lambda_0$; H_1 is $\lambda = \lambda_1 < \lambda_0$

$$r(\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} = \left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\{(\lambda_0 - \lambda_1)\sum x_j\},$$

which is increasing in $\sum x_j$. So the test is based on $\sum x_j$ or $2\lambda_0 \sum X_j$, which is χ^2_{2n} under H_0, and the critical regions are $\{\mathbf{x} : \sum x_j > \frac{1}{2}\lambda_0^{-1}\chi^2_{2n}(\alpha)\}$, where $\chi^2_{2n}(\alpha)$ is the upper α point of χ^2_{2n}. The power is

$$\begin{aligned}\mathbb{P}(2\lambda_0 \sum X_j > \chi^2_\alpha | \, H_1) &= \mathbb{P}\left(2\lambda_1 \sum X_j > \frac{\lambda_1}{\lambda_0}\chi^2_{2n}(\alpha) | \, H_1\right) \\ &= Q_{2n}\left(\frac{\lambda_1}{\lambda_0}\chi^2_{2n}(\alpha)\right)\end{aligned}$$

where Q_{2n} is $1-$ distribution function for χ^2_{2n}.

For comparison, we might base a test on $x_{(1)}$, which has distribution function $1-e^{-n\lambda y}$; size α critical regions are given by $\{\mathbf{x} : x_{(1)} > -(n\lambda_0)^{-1}\ln\alpha\}$, and the power is $\alpha^{\lambda_1/\lambda_0}$, which is $< Q_{2n}\left(\frac{\lambda_1}{\lambda_0}\chi^2_\alpha\right)$ for $n > 1$ and $\lambda_1 < \lambda_0$, and does not depend on n.

3.11 Designing the best test: simple null and composite alternative hypotheses

Suppose now there is a parametric family $\{p(\mathbf{x}| \, \theta) : \theta \in \Theta_1\}$ of alternative p.d.f.'s for $\mathbf{X}$. The power of a size-α critical region R_α generalizes to the size-α **power function**

$$\begin{aligned} pow(\theta; \alpha) &= \mathbb{P}(\mathbf{X} \in R_\alpha | \, \theta) \\ &= \int_{R_\alpha} p(\mathbf{x}| \, \theta) dy \quad \left(\text{or } \sum_{R_\alpha} p(\mathbf{x}| \, \theta) dy\right) \quad (\theta \in \Theta_1).\end{aligned}$$

A size-α critical region R_α is then **uniformly most powerful size** α (UMP size α) if it has maximum power uniformly over Θ_1. A test is UMP if all its critical regions are UMP. More formally

Definition 3.2. *A* **uniformly most powerful** *or UMP test, $\phi_0(\mathbf{X})$, of size α is a test $t(\mathbf{x})$ for which*

(i) $\mathbb{E}_\theta \phi_0(\mathbf{X}) \leq \alpha \quad \forall \, \theta \in \Theta_0$;

(ii) given any other test $\phi(\cdot)$ for which $\mathbb{E}_\theta \phi(\mathbf{X}) \leq \alpha \quad \forall \quad \theta \in \Theta_0$, we have $\mathbb{E}_\theta \phi_0(\mathbf{X}) \geq \mathbb{E}_\theta \phi(\mathbf{X}) \quad \forall \quad \theta \in \Theta_1$.

Such tests cannot be found in general, as this requires that the Neyman-Pearson test should be the same for every pair of simple hypotheses. However, for one sided testing problems, i.e., tests of the form $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, there are a wide class of parametric families for which UMP tests exist. These are distributions that have **monotone likelihood ratio** or MLR.

Definition 3.3. *The family of densities $\{p(\mathbf{x}|\theta), \theta \in \Omega_\theta \subseteq \mathbb{R}\}$ with real scalar parameter θ is said to be of* **monotone likelihood ratio** *if there exists a function $s(\mathbf{x})$ such that the likelihood ratio*

$$\frac{p(\mathbf{x}|\theta_2)}{p(\mathbf{x}|\theta_1)}$$

is a non-decreasing function of $s(\mathbf{x})$ whenever $\theta_1 < \theta_2$.

Note that the same result applies for a non-increasing test statistic, by replacing $t(\mathbf{x})$ by $-t(\mathbf{x})$.

Theorem 3.2. *Suppose $\mathbf{X}$ has a distribution from a family that is monotone likelihood ratio with respect to some continuous test statistic $s(\mathbf{X})$ and we wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$, then a UMP test exists with critical region of the form $s \geq s_\alpha$.*

Proof. For testing $\theta = \theta_0$ against $\theta = \theta_1$ for any specific $\theta_1 \in \Theta_1$, the Neyman-Pearson lemma tells us that the most powerful critical region is given by the likelihood ratio critical region. The LR is a non-decreasing function of $s(\mathbf{y})$ for any $\theta_1 > \theta_0$, and so the critical region is of the form $s \geq s_\alpha$. s_α is determined by the size of the test and depends only on θ_0. Hence, this critical region is identical for all $\theta_1 \geq \theta_0$ and this test is UMP. □

Corollary 3.1. *If $X_1, \ \dots \ , X_n$ are i.i.d with p.d.f. of the form*

$$p(x|\,\theta) = \exp\{a(x)b(\theta) + c(\theta) + d(x)\}$$

with θ a scalar parameter and $b(\theta)$ strictly increasing, then for testing the null hypothesis that $\theta = \theta_0$ against $\theta > \theta_0$ the LR test has critical regions corresponding to large values of $s = \sum a(x_j)$ and is UMP.

Proof For any $\theta_1 > \theta_0$, the LR is

$$\frac{p_{\mathbf{X}}(\mathbf{x}|\,\theta_1)}{p_{\mathbf{X}}(\mathbf{x}|\,\theta_0)} = \exp[\{b(\theta_1) - b(\theta_0)\}s + n\{c(\theta_1) - c(\theta_0)\}].$$

Since $b(\theta_1) > b(\theta_0)$, this is monotone likelihood ratio and so the conditions of Theorem 3.2 are satisfied. This applies to all one-parameter exponential families, e.g. Normal, Binomial, Poisson. There are similar results for $\theta < \theta_0$, when $b(\theta)$ is a decreasing function.

Example.

- Angles : take H_0 to be that angles $X_1, \ \dots \ , X_n$ are i.i.d. and Uniform on $[0, 2\pi)$.

  A set of alternatives representing a type of symmetrical clustering about $y = 0$ has the X_j i.i.d. with von Mises p.d.f.

  $$\frac{\exp(\theta \cos x)}{2\pi I_0(\theta)} \quad (0 \leq x < 2\pi;\ \theta > 0).$$

  So we test the hypothesis $H_0 : \theta = 0$ against the alternative $\theta > 0$.

3.12 Designing the best test: composite null and alternative hypotheses

3.12.1 One-sided tests

Previously we considered tests of hypotheses where the null hypothesis was simple. Testing composite hypotheses is more complex in general. However, the above result for monotone likelihood ratio distributions also applies to one-sided tests of the form $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.

Theorem 3.3. *Suppose* $\mathbf{X}$ *has a distribution from a family that is monotone likelihood ratio with respect to some continuous test statistic* $s(\mathbf{X})$ *and we wish to test* $H_0 : \theta \leq \theta_0$ *against* $H_1 : \theta > \theta_0$*, then*

(a) The test

$$\phi_0(\mathbf{x}) = \begin{cases} 1 & \textit{if } s(\mathbf{x}) > s_0, \\ 0 & \textit{if } s(\mathbf{x}) \leq s_0, \end{cases} \tag{56}$$

is UMP among all tests of size $\leq \mathbb{E}_{\theta_0} \{\phi_0(\mathbf{X})\}$*.*

(b) Given some $0 < \alpha \leq 1$*, there exists an* s_0 *such that the tests in (a) has size exactly equal to* α*.*

Proof. 1. From Theorem 3.2, ϕ_0 is UMP for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

2. $\mathbb{E}_\theta\{\phi_0(\mathbf{x})\}$ is a non-decreasing function of θ. If we have $\theta_2 < \theta_1$ and $\mathbb{E}_{\theta_2}\{\phi_0(\mathbf{x})\} = \beta$, then the trivial test $\phi(\mathbf{x}) = \beta$ has $\mathbb{E}_{\theta_1}\{\phi(\mathbf{x})\} = \beta$. The test ϕ_0 is UMP for testing θ_2 against θ_1 and so it must be at least as good as ϕ, i.e., $\mathbb{E}_{\theta_1}\{\phi_0(\mathbf{x})\} \geq \beta$. Hence, if we construct the test with $\mathbb{E}_{\theta_0}\{\phi_0(\mathbf{x})\} = \alpha$, then $\mathbb{E}_\theta\{\phi_0(\mathbf{x})\} \leq \alpha$ for all $\theta \leq \theta_0$, so ϕ_0 is also of size α under the larger hypothesis $H_0 : \theta \leq \theta_0$.

3. For any other test ϕ that is of size α under H_0, we have $\mathbb{E}_{\theta_0}\{\phi(\mathbf{x})\} \leq \alpha$ and by the Neyman-Pearson lemma $\mathbb{E}_{\theta_1}\{\phi_(\mathbf{x})\} \leq \mathbb{E}_{\theta_1}\{\phi_0(\mathbf{x})\}$ for any $\theta_1 > \theta_0$. This shows that this test is UMP among all tests of its size.

4. If α is specified we must show that there exists a s_0 such that $\mathbb{P}_{\theta_0}\{s(\mathbf{X}) > s_0\} = \alpha$, but this follows from the assumption that $s(\mathbf{X})$ is continuous.

□

3.12.2 Two-sided tests

In more general situations we will be interested in testing hypotheses of the form $H_0 : \theta \in \Theta_0$, where Θ_0 is either an interval $[\theta_1, \theta_2]$ for $\theta_1 < \theta_2$ or a single point $\Theta_0 = \{\theta_0\}$, against the generic alternative $H_1 : \theta \in \Theta_1$, with $\Theta_1 = \mathbb{R}/\Theta_0$. For a family with monotone likelihood ratio with respect to a statistic $s(\mathbf{X})$, we might expect a good test to have a test function of the form

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } s(\mathbf{x}) > s_2 \text{ or } s(\mathbf{x}) < s_1, \\ \gamma(\mathbf{x}) & \text{if } s(\mathbf{x}) = s_2 \text{ or } s(\mathbf{x}) = s_1, \\ 0 & \text{if } s_1 < s(\mathbf{x}) < s_2. \end{cases}$$

Such a test is called **a two-sided test**. For such two-sided tests, we cannot usually find a UMP test. However, under certain circumstances it is possible to find a **uniformly most powerful unbiased** (UMPU) test.

Definition 3.4. *A test $\phi(\mathbf{y})$ of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is called* **unbiased of size** *α if*

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta \left\{ \phi(\mathbf{Y}) \right\} \leq \alpha$$

and

$$\mathbb{E}_\theta \left\{ \phi(\mathbf{Y}) \right\} \geq \alpha \text{ for all } \theta \in \Theta_1.$$

In other words, an unbiased test is one which has higher probability of rejecting H_0 when it is false than when it is true. Note that if the power function is a continuous function of θ then an unbiased test of size α must have size equal to α on the boundary of the critical region (since the size is less than or equal to α within the critical region and greater than or equal to α outside).

Definition 3.5. *A test which is uniformly most powerful among the set of all unbiased tests is called* **uniformly most powerful unbiased***.*

For a scalar exponential family of the form given in Corollary 3.1 the following theorem holds

Theorem 3.4. *If $X_1, \ \ldots \ , X_n$ are i.i.d with p.d.f. of the form*

$$p(x|\,\theta) = \exp\{a(x)b(\theta) + c(\theta) + d(x)\}$$

with θ a scalar parameter and $b(\theta)$ strictly increasing, then there exists a unique UMPU test of size α, ϕ', for testing the hypothesis $H_0 : \theta \in [\theta_1, \theta_2]$, against the generic alternative $H_1 : \theta \in \mathbb{R} - [\theta_1, \theta_2]$, of the form

$$\phi'(\mathbf{x}) = \begin{cases} 1 & \text{if } s(\mathbf{x}) > s_2 \text{ or } s(\mathbf{x}) < s_1, \\ \gamma_j & \text{if } s(\mathbf{x}) = s_j, \\ 0 & \text{if } s_1 < s(\mathbf{x}) < s_2. \end{cases} \tag{57}$$

where $S = \sum a(x_j)$, for which

$$\mathbb{E}_{\theta_j} \phi'(\mathbf{X}) = \mathbb{E}_{\theta_j} \phi(\mathbf{X}) = \alpha, \qquad j = 1, 2.$$

The boundaries of the critical region, s_1, s_2, and the rejection probabilities on the boundaries, γ_1, γ_2, are determined from the conditions $\mathbb{E}_{\theta_j} \phi'(\mathbf{X}) = \alpha$.

Example. Suppose a sample Y is drawn from an Exp(λ) distribution, so that $f(y|\lambda) = \lambda \exp(-\lambda y)$. Construct a uniformly most powerful unbiased test of size $\alpha = 0.05$ of the hypothesis $H_0 : \lambda \in [1, 2]$ against the generic alternative $\lambda \in [0, 1) \cup (2, \infty)$.

For a single sample from the exponential distribution, the sufficient statistic is the observed value, y. Using the previous result, the UMPU test is of the form (57). The probability that $s = s_i$ is zero for any single value s_i and therefore the γ_i's do not need to be determined. The boundaries of the critical region can be found from the constraints

$$\alpha = 0.05 = 1 - \exp(-s_1) + \exp(-s_2) = 1 - \exp(-2s_1) + \exp(-2s_2),$$

from which we find $s_1 = 0.02532$ and $s_2 = 3.6889$. The corresponding power function $\eta(\lambda)$ is shown in Figure 2. This shows that the test is unbiased as the probability of rejecting H_0 is less than or equal to the size α within the region defined by H_0, it is equal to α on the boundary, and greater than α everywhere outside that region.

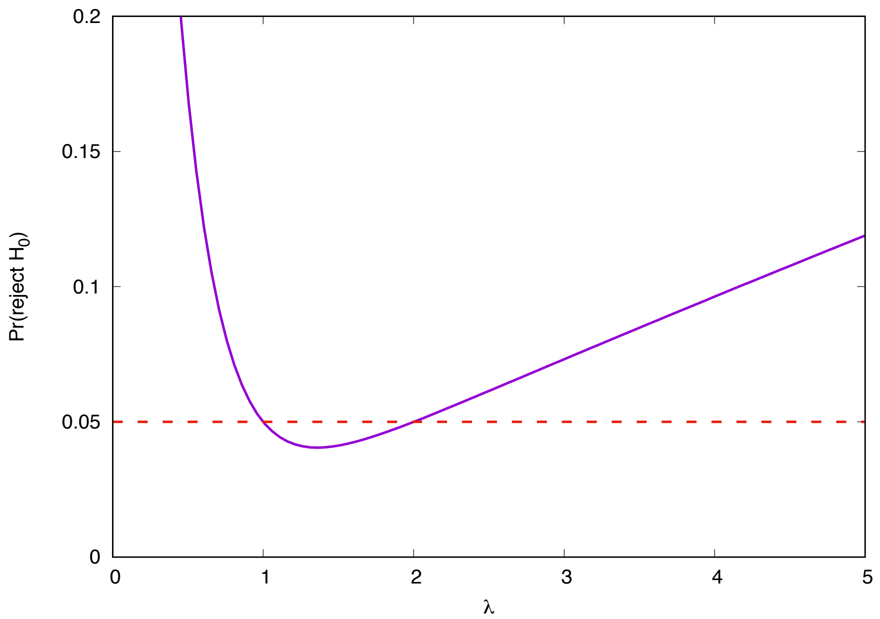


Figure 2: Power of the UMPU test of $\lambda \in [1,2]$ against a generic alternative for an exponential distribution, as a function of λ, i.e., $\mathbb{P}_\lambda(\text{reject } H_0)$. The horizontal line indicates the size of the test, $\alpha = 0.05$.

3.12.3 Testing a point null hypothesis

A test of the null hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ can be considered as the limit of the preceding two-sided test when $\theta_2 - \theta_1 \to 0$. Therefore, as a corollary to the previous result, there must exist a unique UMPU test, ϕ', of this hypothesis of the form (57) for which

$$\mathbb{E}_{\theta_0}\{\phi'(X)\} = \alpha, \qquad \frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta\{\phi'(X)\}|_{\theta=\theta_0} = 0. \tag{58}$$

Differentiability of the power function for any test function is ensured from the assumption that the distribution is in the exponential family.

Example. Returning to the example of the preceding section of a single sample from an $\text{Exp}(\lambda)$ distribution, if we instead want to test the hypothesis that $\lambda = 1$ then we proceed as before, but the constraints on the boundary of the rejection region are now

$$\begin{aligned} \alpha &= 0.05 = 1 - \exp(-t_1) + \exp(-t_2), \\ 0 &= t_1 \exp(-t_1) - t_2 \exp(-t_2), \end{aligned}$$

which can be solved numerically to give $t_1 = 0.0423633$, $t_2 = 4.76517$. The power function is shown in Figure 3. We see that it reaches a minimum of $\alpha = 0.05$ at $\theta = \theta_0$ so it is unbiased and of size α as desired.

3.13 Designing the best test: similar Tests

So far we have focussed on tests of one-parameter distributions. However, often the distribution will depend on more than one parameter. In that case we are interested in tests

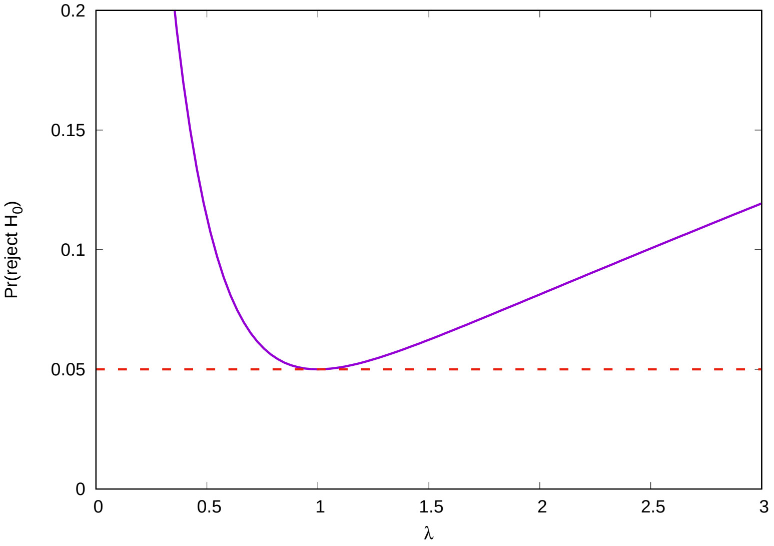


Figure 3: Power of the UMPU test of $\lambda = 1$ against a generic alternative for an exponential distribution, as a function of λ, i.e., $\mathbb{P}_\lambda(\text{reject } H_0)$. The horizontal line indicates the size of the test, $\alpha = 0.05$.

that perform as well as possible in inferring the value of one parameter of the distribution, irrespective of the value of the other parameters of the distribution. This gives rise to the notion of a **similar** test.

Definition 3.6. *Suppose $\theta = (\psi, \lambda)$ and the parameter space is of the form $\Omega_\theta = \Omega_\psi \times \Omega_\lambda$. Suppose we wish to test the null hypothesis $H_0 : \psi = \psi_0$ against the alternative $H_1 : \psi \neq \psi_0$, with λ treated as a nuisance parameter. Suppose $\phi(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$ is a test of size α for which*

$$\mathbb{E}_{\psi_0,\lambda}\left\{\phi(\mathbf{x})\right\} = \alpha \text{ for all } \lambda \in \Omega_\lambda.$$

Then ϕ is called a **similar test of size** *α.*

This definition can be extended to composite null hypotheses. If the null hypothesis is of the form $\theta \in \Theta_0$, where Θ_0 is a subset of Ω_θ, then a similar test is one for which $\mathbb{E}_\theta\left\{\phi(\mathbf{x})\right\} = \alpha$ on the boundary of Θ_0.

If a test is uniformly most powerful among all similar tests then it is called **UMP similar**. There is close connection to UMPU tests. If the power function of a test is continuous then we saw earlier that any unbiased test of size α must have size exactly equal to α on the boundary, i.e., it must be similar. In such cases, if we can find a UMP similar test and it turns out to also be unbiased, then it is necessarily UMPU.

Moreover, in many cases it is possible to demonstrate that a test which is UMP among all tests based on the conditional distribution of a statistic S given the value of an ancillary statistic A, this test is UMP among all similar tests. In particular, this applies if A is a complete sufficient statistic for the variables λ.

One common situation in which this occurs is for multi-parameter exponential families, for which the likelihood can be written

$$p(x|\theta) = \exp\left\{\sum_{i=1}^{p} A_i(x)B_i(\theta) + C(\theta) + D(x)\right\}.$$

Consider a test of the form $H_0 : B_1(\theta) \leq \theta_1^*$ against $H_1 : B_1(\theta) > \theta_1^*$. If we take $s(\mathbf{x}) = \sum_j A_1(x_j)$ and $A = (\sum_j A_2(x_j), \ldots, \sum_j A_p(x_j))$, then the conditional distribution of S given A is also of the exponential form and doesn't depend on $B_2(\theta), \ldots, B_p(\theta)$, so A is both sufficient and complete for $B_2(\theta), \ldots, B_p(\theta)$. The Conditionality Principle suggests we should make inference about $B_1(\theta)$ based on the conditional distribution of S given A. Tests constructed in this way are UMPU (Ferguson 1967). The optimal one-sided test is then of the following form. Based on observations $s_1 = \sum_j A_1(x_j)$, $s_2 = \sum_j A_2(x_j), \ldots, s_p = \sum_j A_p(x_j)$, we reject H_0 if and only if $s_1 > s_1^*$, where s_1^* is calculated from

$$\mathbb{P}_{B_1(\theta)=\theta_1^*}\left\{S_1 > s_1^* | S_2 = s_2, \ldots, S_p = s_p\right\} = \alpha.$$

It can be shown this is a UMPU test of size α.

Similarly, to construct a two-sided test of $H_0 : \theta_1^* \leq B_1(\theta) \leq \theta_1^{**}$ against $B_1(\theta) < \theta_1^*$ or $B_1(\theta) > \theta_1^{**}$, we first define the conditional power function

$$w_{\theta_1}(\phi|s_2, \ldots, s_p) = \mathbb{E}_{\theta^1}\left\{\phi(S_1)|S_2 = s_2, \ldots, S_p = s_p\right\}.$$

Then we can construct a two-sided conditional test of the form

$$\phi'(s_1) = \begin{cases} 1 & \text{if } s_s < s_1^* \text{ or } s_1 > s_1^{**}, \\ 0 & \text{if } s_1^* \leq s_1 \leq s_1^{**}, \end{cases}$$

where s_1^* and s_1^{**} are chosen such that

$$w_{\theta_1}(\phi'|s_2, \ldots, s_p) = \alpha \quad \text{when } B(\theta_1) = \theta_1^* \text{ or } B(\theta_1) = \theta_1^{**}.$$

It can be shown that these tests are also UMPU of size α. If the test is of a simple hypothesis $B(\theta_1) = \theta_1^*$ against the generic alternative $B(\theta_1) \neq \theta_1^*$ then the test is of the same form but the conditions are that the power function is equal to α and its derivative with respect to θ is equal to 0, as in Eq. (58).

3.14 Generalized likelihood ratio tests

In the previous sections we focussed on finding the "best" tests by one metric or another. However, as we have seen this is not always easy and the resulting test statistics are not always straightforward to evaluate. Under many circumstances, in the limit $n \to \infty$, the likelihood ratio follows a χ^2 distribution and so this can be used to construct a test that is valid asymptotically.

In particular, suppose we are testing $H_0 : \vec{\theta} \in \Theta_0$ versus $H_1 : \vec{\theta} \in \Theta_1$. We define the likelihood ratio

$$L_X(H_0, H_1) = \frac{\sup_{\vec{\theta}\in\Theta_1} p(x|\theta)}{\sup_{\vec{\theta}\in\Theta_0} p(x|\theta)}$$

and denote by $p = |\Theta_1 - \Theta_0|$ the difference in the numbers of degrees of freedom in the unknown parameters between the two hypotheses. Then as $n \to \infty$

$$2\log L_X(H_0, H_1) \sim \chi_p^2$$

under H_0 and tends to be larger under H_1. Therefore critical regions of the form $2 \log L_X > \chi_p^2(\alpha)$ give tests of approximately size α.

The interpretation of p is the number of constraints that have been placed to reduce the, typically more general, alternative hypothesis, to the more restrictive null hypothesis. For example, the null hypothesis might be specified by fixing the values of p of the parameters, or by imposing p linear constraints on the parameters, or by writing the k parameters of Θ_1 as functions of an alternative $k - p$ dimensional parameter space.

4 Bayesian Theory

As we have seen, in frequentist statistics statements are made with reference to repetitions of the same experiment with parameters fixed. In Bayesian statistics, parameters are no longer regarded as fixed, but are themselves random variables. The probability distribution of the parameter values before taking data, the **prior distribution**, is updated to a probability distribution after taking data, the **posterior distribution**, through the likelihood of the observed data. This update is achieved through **Bayes' Theorem**. Bayesian inference attempts to say as much as possible about the unknown parameter distribution based on the observed data only, without reference to future repetitions of the same experiment. Bayesian posteriors are probability distributions on the unknown parameter and can be interpreted and manipulated in that way, as statements about the relative probability that the parameter takes different values.

The derivation of Bayes' theorem is a mathematical result that follows from the definition of conditional probability, as we will see below, but it is how this result is applied to interpret data, and the philosophical distinction in the interpretation of the parameter values that distinguishes the frequentist and Bayesian approach. Typically, in any given observation, the actual parameter values that led to the generation of the observed data are fixed, not random, but the Bayesian interpretation is that you can never by sure of what the unknown parameter is, and so it is appropriate to consider it to be a random variable. In many cases you will not be able to repeat a particular experiment. Gravitational wave observations are a good example of this — we cannot choose what events occur in the Universe, so every observed event is a unique, non-repeatable, experiment. In such contexts, the frequentist approach of referencing theoretical repetitions cannot really be seen as representative of reality. In cases where it is possible to repeat an experiment with the unknown parameters fixed, the Bayesian posterior converges to the true parameter value asymptotically and so can still be used to represent the current level of uncertainty in the parameter.

Frequentist concepts such as significance and hypothesis testing have been incorporated into the Bayesian framework, but the interpretation in the latter context is not always clean. It is therefore useful to have familiarity with both sets of tools to be fully quipped to handle any kind of data analysis problem.

4.1 Conditional probability

It is often the case that a process generates more than one potentially measurable random output, but only a subset of these are measurable. If the variables are independent then measuring one would not provide any information about the others, but when there are inter-dependencies the observation of a random variable can provide information about other variables with which it is correlated. For example, suppose we have a bag containing 100 balsa, of which 10 are red and stripy, 20 are blue and stripy, 30 are red and spotted and 40 are blue and spotted. In total there are 30 stripy balls out of the 100 and therefore the probability that a randomly chosen ball is stripy is 3/10. However, out of the 40 red balls there are only 10 that are stripy, and so if we have observed that the ball is red the probability that it is also stripy is now 1/4.

The **conditional probability** of an event A, given some other event B is defined as

$$p(A|B) = \frac{p(A \cap B)}{B}.$$

In other words, this is the fraction that both A and B occur, our of all the times that B occurs. This can be rewritten in two different ways by interchanging A and B

$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A).$$

Rearranging this identity we obtain **Bayes' Theorem**

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

4.2 Bayesian inference

Bayes' Theorem is a mathematical identity, but it becomes philosophically distinct from frequentist approaches when it is applied to inference. In Bayesian inference, the event A is taken to be an observation of data, $\mathbf{x}$, and the event B is taken to be the value of some unknown parameters, $\vec{\theta}$, characterising the system being observed. Bayes' Theorem becomes

$$p(\vec{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\vec{\theta})p(\vec{\theta})}{p(\mathbf{x})}.$$

In this context $p(\mathbf{x}|\vec{\theta})$ is the likelihood (the same function of data and parameters as in the frequentist case), $p(\vec{\theta})$ is the **prior** distribution of source parameter values, $p(\vec{\theta}|\mathbf{x})$ is the **posterior** distribution on the source parameter values and $p(\mathbf{x})$ is the **evidence** for the model under consideration. In a parameter estimation context, the evidence, which does not depend on parameter values, is a normalisation constant that can be ignored. However, it plays an important role in Bayesian hypothesis testing, which will be discussed in section 4.6.

Example: Medical testing We suppose that a medical test for a disease is 95% effective but has a 1% false alarm rate and the prevalence of the disease in the population is 0.5%. You test positive for the disease. What is the probability you do in fact have it?

The term "95% effective" means that if you have the disease the test gives a positive result 95% of the time. The term 1% false alarm rate means that if you do not have the disease you test positive 1% of the time. We can now apply Bayes theorem with data $\mathbf{x} =$ 'positive test' and parameter $\theta =$'disease status' taking values 'infected' or 'not infected'. The likelihood is

$$p(\text{positive}|\text{infected}) = 0.95, \qquad p(\text{positive}|\text{not infected}) = 0.01.$$

The prior is based on the known prevalence in the population

$$p(\text{infected}) = 1 - p(\text{not infected}) = 0.005.$$

The posterior is then

$$\begin{aligned} p(\text{infected}|\text{positive}) &= \frac{p(\text{positive}|\text{infected})p(\text{infected})}{p(\text{positive}|\text{infected})p(\text{infected}) + p(\text{positive}|\text{not infected})p(\text{not infected})} \\ &= \frac{0.95 * 0.005}{0.95 * 0.005 + 0.01 * 0.995} = 0.323. \qquad (59) \end{aligned}$$

So you are more likely not to be infected than to be infected if you get a positive test result. The solution is to get a second opinion. If you take a second (independent) test and it is also positive your posterior probably of being infected is now

$$p(\text{infected}|\text{2nd positive}) = \frac{0.95 * 0.323}{0.95 * 0.323 + 0.01 * 0.677} = 0.978 = \frac{0.95^2 * 0.005}{0.95^2 * 0.005 + 0.01^2 * 0.995}.$$

The first of these two results follows from using the posterior from the first test as a prior for the second. The second result follows from regarding the observed data as "two independent positive tests".

Example: Blood evidence Based on other evidence, a detective is 50% sure that a particular suspect has committed a murder. Then new evidence comes to light. A small amount of blood, of type B, is found at the scene. This is not the victim's blood type, but it is the blood type of the suspect. Such a blood type has a prevalence of 2% in the population. What is the detective's confidence in the guilt of the suspect in light of this new evidence?

The likelihood is

$$p(\text{type B blood}|\text{guilty}) = 1, \qquad p(\text{type B blood}|\text{not guilty}) = 0.02.$$

The prior is $p(\text{guilty}) = 0.5$ and so the posterior is

$$\begin{aligned} p(\text{guilty}|\text{type B blood}) &= \frac{p(\text{type B blood}|\text{guilty})p(\text{guilty})}{p(\text{type B blood}|\text{guilty})p(\text{guilty}) + p(\text{type B blood}|\text{not guilty})p(\text{not guilty})} \\ &= \frac{0.5}{0.5 + 0.01} = 0.98. \end{aligned} \tag{60}$$

4.3 Choice of prior

The prior plays a key role in Bayesian parameter inference. It expresses the current state of our understanding about parameter values, and it is updated to the posterior using data via the likelihood. Mathematically, the prior represents the distribution of the unknown parameter value in nature, but usually this is not known. In that case, the prior reflects the current state of knowledge about the parameter values, which may come from previous experiments or expert opinion or not be known.

4.3.1 Informative/expert priors

If information is available, it is appropriate to use informative priors. For example, if previous measurements have been made of a quantity it is reasonable to use the posterior from those measurements as a prior for the next measurement, as we saw in the medical test example above. Alternatively, even if a measurement has not been made directly, "experts" may be able to give a reasonable range or distribution for the parameter based on experience in other situations. One criticism that is often levelled at Bayesian inference is that the result can depend on the assumed prior. However, the Bayesian response is that this is desired behaviour — if we have additional information from prior knowledge, then it is the correct thing to do to include that in our conclusions based on subsequent observed data.

The process of constructing a prior based on the opinion of experts is known as **elicitation**. Sometimes, elicitation may result in different priors from different experts. In that case a **mixture prior** can be constructed

$$p(\vec{\theta}) = \sum_{j=1}^{J} \omega_j p_j(\vec{\theta})$$

where j labels which of the J experts we are referring to, $p_j(\vec{\theta})$ is the prior elicited from that expert, and ω_j is the weight given to that expert (or set of experts).

If the prior is based on the posterior from previous observations it is normally clear how to fold this in. If the prior comes from expert opinion, it may be possible to use this in several different ways. In that case, care must be taken to be as conservative as is reasonably possible in the use of that prior information, to avoid making conclusions form the data that are too strong.

4.3.2 Conjugate priors

It is convenient to choose a form for the prior that ensures the posterior takes the same form. In such situations, the posterior from an experiment can be directly be used as a prior for the next experiment and so on. Such a prior is called **conjugate**.

Definition: A family of distributions, $\mathcal{F}$, is **conjugate** to a family of sampling distributions, $\mathcal{P}$, if, whenever the prior belongs to the family $\mathcal{F}$, the posterior belongs to the same family, for any number and value of observations from $\mathcal{P}$.

The form of the conjugate prior depends on the nature of the probability distribution, $\mathcal{P}$, from which the observed data is drawn. This gives rise to a number of conjugate families. In particular, any distribution in the exponential family

$$p(x|\,\theta) = \exp\left\{ \sum_{j=1}^{K} A_j(x) B_j(\vec{\theta}) + C(\vec{\theta}) + D(x) \right\} \; \forall x, \vec{\theta}$$

has a conjugate prior in the exponential family of the form

$$p(\vec{\theta}|\vec{\chi}, \nu) = p(\vec{\chi}, \nu) \exp\left[\vec{\theta}^T \vec{\chi} - \nu A(\vec{\theta}) \right] \tag{61}$$

where ν and $\vec{\chi}$ are the hyperparameters of the prior distribution.

A full list of conjugate priors can be found in the conjugate prior entry on wikipedia, but the three most widely used are the Beta-Binomial, Poisson-Gamma and Normal-Normal families, and we will discuss these further here.

Beta-Binomial model Suppose our observed data $\mathbf{X} \sim$Bin(n,p) with likelihood

$$p(\mathbf{x}|p) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}.$$

The conjugate prior is the Beta(a,b) distribution with density

$$p(p) = \frac{1}{B(a,b)} p^{a-1} (1-p)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1}.$$

Observing binomial distributed data and using the Beta prior gives a posterior

$$\begin{aligned} p(p \mid x) &\propto p(x \mid p)p(p) \\ &= \binom{n}{x} p^x (1-p)^{n-x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \\ &\propto p^{a+x-1}(1-p)^{b+n-x-1}. \end{aligned}$$

So the posterior is also a Beta distribution

$$p(p \mid x) = \mathrm{Beta}(a+x, b+n-x).$$

The mean and variance of a Beta(a, b) distribution are

$$\mathbb{E}(\mathbf{X}) = \frac{a}{a+b}, \qquad \mathrm{var}(\mathbf{X}) = \frac{ab}{(a+b)^2(a+b+1)}.$$

The posterior mean is therefore

$$\mathbb{E}(p|x) = \frac{a+x}{a+b+n}$$

which we compare to the mean in the observed data of x/n. One interpretation of the prior data is that it represents having observed $a-1$ events in $a+b-2$ previous trials. If a and b are kept fixed and $n, x \to \infty$ the posterior mean tends to the maximum likelihood estimator x/n and the posterior variance tends to zero.

Poisson-Gamma model Suppose now that we are observing data, $X_1, \ldots, X_n$, from a Poisson distribution, $\mathbf{X} \sim$Pois(λ), with likelihood

$$p(\mathbf{x} \mid \lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i} \mathrm{e}^{-\lambda}}{x_i!} \right\}.$$

The conjugate prior is the Gamma(m, μ) distribution

$$p(\lambda|m, \mu) = \frac{1}{\Gamma(m)} \mu^m \lambda^{m-1} \mathrm{e}^{-\mu\lambda},$$

which has mean m/μ and variance m/μ^2. With this prior the posterior is

$$\begin{aligned} p(\lambda \mid \mathbf{x}) &\propto p(\mathbf{x}|\lambda)p(\lambda) \\ &= \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i} \mathrm{e}^{-\lambda}}{x_i!} \right\} \frac{1}{\Gamma(m)} \mu^m \lambda^{m-1} \mathrm{e}^{-\mu\lambda} \\ &\propto \mathrm{e}^{-n\lambda - \mu\lambda} \lambda^{\sum_{i=1}^{n} x_i + m - 1} \\ &\propto \mathrm{Gamma}(m + n\bar{x}, \mu + n). \end{aligned} \tag{62}$$

The posterior mean can be seen to equal

$$\mathbb{E}(p(\lambda \mid \mathbf{x})) = \frac{m + n\bar{x}}{m+n} = \bar{x} \left(\frac{n}{n+m} \right) + \frac{m}{\mu} \left(1 - \frac{n}{n+m} \right),$$

i.e., it is a compromise between the prior mean, m/μ, and the maximum likelihood estimator $\bar{x}$. As the number of samples increases, more weight is placed on the data and less on the prior, as expected.

Normal-Normal/Normal-Gamma model Now we consider $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, and likelihood

$$p(\mathbf{x}|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right].$$

We assume first that σ^2 is known. The conjugate prior in this case is the Normal distribution, $N(\mu_0, \sigma_0^2)$,

$$p(\mu \mid \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right].$$

The posterior is

$$\begin{aligned} p(\mu \mid \mathbf{x}, \sigma^2) &\propto p(\mathbf{x} \mid \mu, \sigma^2) p(\mu|\mu_0, \sigma_0^2) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right\} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2\sigma_0^2}\left[\mu^2(n\sigma_0^2 + \sigma^2) - 2\mu(n\bar{y}\sigma_0^2 + \mu_0\sigma^2)\right]\right\}, \end{aligned}$$

which can be recognized as a $N(\mu_n, \sigma_n^2)$ distribution, where

$$\mu_n = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \qquad \sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}. \tag{63}$$

Writing these results in terms of $\tau = 1/\sigma^2$, which is called the **precision** of the Normal distribution we can see

$$\mu_n = \frac{\tau_0}{\tau_0 + n\tau}\mu_0 + \frac{n\tau}{\tau_0 + n\tau}\bar{y}$$

so once again the posterior mean is a balance between the prior mean and the sample mean, with the relative weighting determined by both the number of observations and the relative precision of the observations and the prior.

If we suppose that μ is known (which is an unrealistic assumption in practice), but the variance is uncertain, then we can obtain a conjugate prior by using a Gamma(a, b) prior on the precision

$$p(\tau|a, b) \propto \tau^{a-1}e^{-b\tau}$$

and obtain the posterior

$$\begin{aligned} p(\tau \mid \mathbf{x}, \mu) &\propto p(\mathbf{x} \mid \mu, \tau) p(\tau|a, b) \\ &\propto \tau^{n/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \tau^{a-1}e^{-b\tau} \\ &= \tau^{a+n/2-1} \exp\left\{-\tau\left(b + \frac{1}{2}\sum_i (x_i - \mu)^2\right)\right\} \\ &\sim \text{Gamma}\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right). \end{aligned}$$

It is common practice to take the limit in which a and b are both very small and then the posterior becomes

$$p(\tau \mid \mathbf{x}, \mu) = \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \quad \Rightarrow \quad \mathbb{E}\left[\tau \mid \mathbf{x}, \mu\right] = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right)^{-1},$$

so the posterior expectation of the precision is approximately the same as the (frequentist) sample precision (up to a factor of $n/(n-1)$).

Finally we assume that both μ and σ^2 are unknown. It would be reasonable to just multiply together the two previous priors, but this does not result in a conjugate prior, essentially because the posterior on μ in the first case depends on the known variance σ^2. However, we can find a correlated conjugate prior (writing $\tau = 1/\sigma^2$ as before) by writing

$$\mu \sim N(\mu_0, 1/(n_0\tau)), \quad \tau \sim \text{Gamma}(a, b),$$

or, explicitly,

$$p(\mu, \tau|\mu_0, n_0, a, b) \propto \left(\frac{n_0\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left[-\frac{n_0\tau}{2}(\mu-\mu_0)^2\right]\tau^{a-1}e^{-b\tau}.$$

The posterior on μ, conditioned on τ, $p(\mu|\tau, \mathbf{x})$, is given by the same expression as before

$$p(\mu|\tau, \mathbf{x}) \sim N\left(\frac{n_0\mu_0 + n\bar{x}}{n_0+n}, \frac{1}{(n_0+n)\tau}\right).$$

The posterior on τ can be found by considering the combined posterior, being careful not to drop any terms that depend on μ or τ

$$\begin{aligned} p(\mu, \tau|\mathbf{x}) &\propto \sqrt{\tau}\exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right]\tau^{\frac{n}{2}}\exp\left[-\frac{n_0\tau}{2}(\mu-\mu_0)^2\right]\tau^{a-1}e^{-b\tau} \\ &= \tau^{a+\frac{n}{2}-1}\exp\left[-\left(b - \frac{(n\bar{x}+n_0\mu_0)^2}{2(n+n_0)} + \frac{1}{2}n_0\mu_0^2 + \frac{1}{2}\sum x_i^2\right)\tau\right] \times \\ &\quad \times \left(\sqrt{\frac{(n+n_0)\tau}{2\pi}}\exp\left[-\frac{(n+n_0)\tau}{2}\left(\mu - \frac{(n\bar{x}+n_0\mu_0)}{n+n_0}\right)^2\right]\right). \end{aligned} \tag{64}$$

If we now marginalise over μ, the round bracketed term on the final line integrates to a constant, independent of τ, and the term inside the exponent on the penultimate line can be simplified to obtain

$$\begin{aligned} p(\tau|\mathbf{x}) &\propto \tau^{a+\frac{n}{2}-1}\exp\left[-\left(b + \frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{nn_0}{2(n+n_0)}(\mu_0-\bar{x})^2\right)\tau\right] \\ &\Rightarrow p(\tau|\mathbf{x}) \sim \text{Gamma}\left(a+\frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{nn_0}{2(n+n_0)}(\mu_0-\bar{x})^2\right). \end{aligned} \tag{65}$$

And so this is also a conjugate prior model, called the Normal-Gamma model.

4.3.3 Using expert information with conjugate priors

If expert prior information is in the form of a posterior from a previous experiment the form of the distribution is fixed. However, in other circumstances it can be possible to express the prior information in the form of a particular choice of parameters for a conjugate prior. This is most clearly seen with an example.

Example: Consider a drug to be given for relief of chronic pain. Experience with similar compounds has suggested that response rates, p, between 0.2 and 0.6 could be feasible. We plan to observe the response rate in n patients and want to infer a posterior on p. Propose a suitable conjugate prior for p based on the available information.

A response rate between 0.2 and 0.6 could be used to set a uniform prior in that range. However, this is not conjugate to the binomial distribution that determines the observed data. Therefore, it would be better to use a conjugate prior. A $U[0.2, 0.6]$ distribution has mean 0.4 and standard deviation of 0.1. We can find a Beta distribution that has the same mean and standard deviation. Rearranging the equations given earlier we deduce $\text{Beta}(a = 9.2, b = 13.8)$ has the desired mean and variance. This prior is conjugate and reflects the expert opinion as regards the expected response rate for the drug. Suppose now we observe $n = 20$ patients and $x = 15$ respond positively. The posterior is then $\text{Beta}(9.2 + 15, 13.8 + 5) = \text{Beta}(24.2, 18.8)$. The prior, (scaled) likelihood and posterior are illustrated in Figure 4.

4.3.4 Mixture priors

The use of a conjugate prior can be somewhat restrictive as there is limited flexibility within the prior family. However, one way to get around this is by using **mixture priors**. A mixture prior is of the form

$$p(\vec{\theta}) = \sum_{j=i}^{J} \pi_j p(\vec{\theta} \mid \vec{\psi}_j), \quad \sum_{j=1}^{J} \pi_j = 1. \tag{66}$$

Here $\{\pi_j\}$ are called the mixture weights and it is assumed that the hyperparameters, ψ_j, are different in each component. If the mixture components are all drawn from the conjugate prior family, then the mixture prior is also conjugate.

Example: Beta-Binomial mixture prior Suppose $X \sim \text{Bin}(n, p)$ and we use a prior on p that is a mixture distribution

$$p(p|a_1, b_1, a_2, b_2) = \pi\text{Beta}(a_1, b_1) + (1 - \pi)\text{Beta}(a_2, b_2).$$

What is the posterior distribution for p?

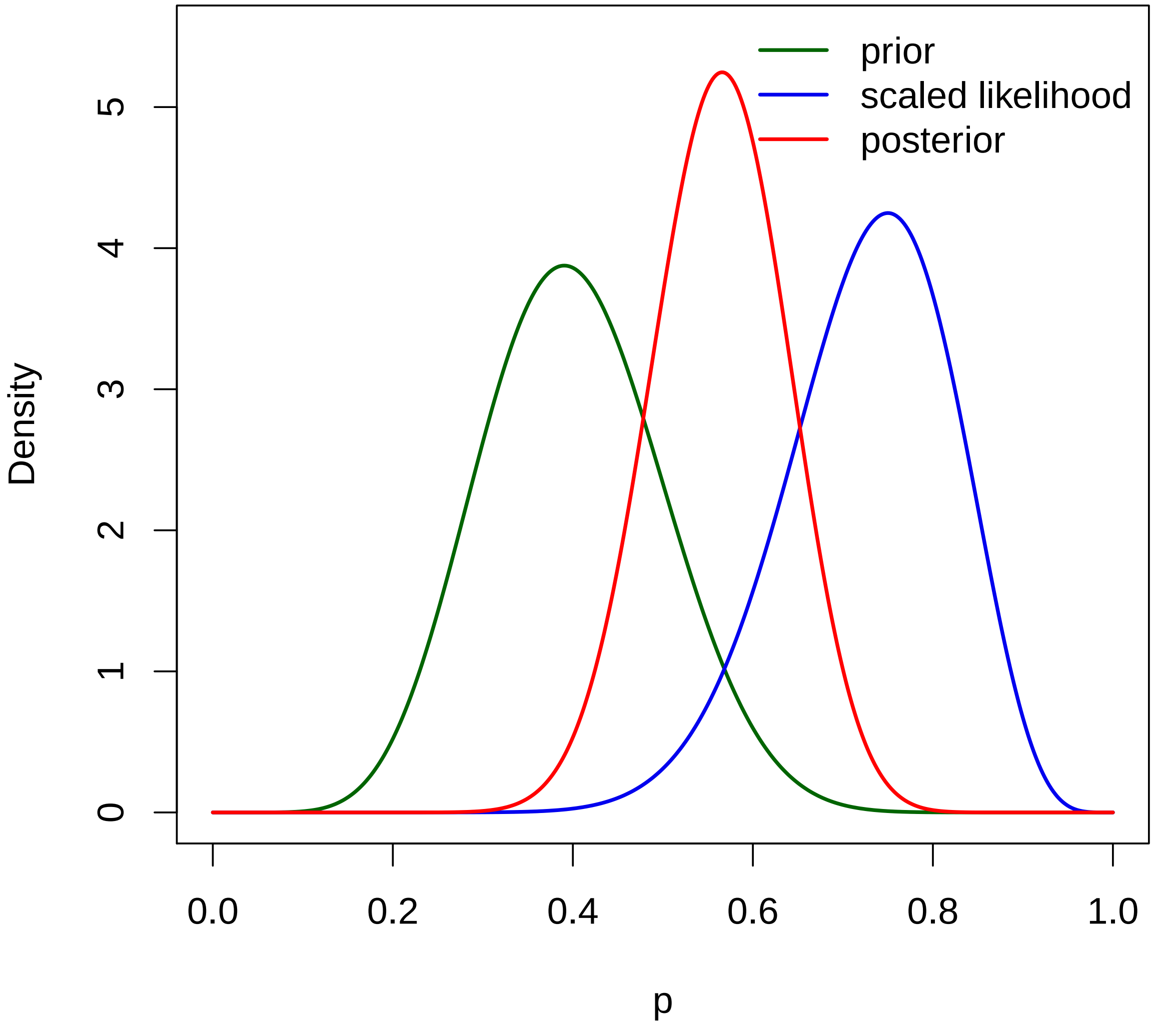


Figure 4: Conjugate prior, Beta$(9.2, 13.8)$, likelihood, Bin$(20, p)$, and posterior, Beta$(24.2, 18.8)$ for the drug response problem described in the text. The likelihood has been rescaled to ensure it has a similar height to the prior and posterior distributions.

Solution: We find the posterior as follows

$$\begin{aligned}
p(p \mid x) &\propto \binom{n}{x} p^x(1-p)^{n-x}\left\{\pi\frac{1}{B(a_1,b_1)}p^{a_1-1}(1-p)^{b_1-1}+(1-\pi)\frac{1}{B(a_2,b_2)}p^{a_2-1}(1-p)^{b_2-1}\right\}\\
&\propto \pi\frac{1}{B(a_1,b_1)}p^{a_1+x-1}(1-p)^{b_1+n-x-1}+(1-\pi)\frac{1}{B(a_2,b_2)}p^{a_2+x-1}(1-p)^{b_2+n-x-1}\\
&=\pi\frac{B(a_1+x,b_1+n-x)}{B(a_1,b_1)}\frac{1}{B(a_1+x,b_1+n-x)}p^{a_1+x-1}(1-p)^{b_1+n-x-1}\\
&\quad+(1-\pi)\frac{B(a_2+x,b_2+n-x)}{B(a_2,b_2)}\frac{1}{B(a_2+x,b_2+n-x)}p^{a_2+x-1}(1-p)^{b_2+n-x-1}\\
&=\pi\frac{B(a_1+x,b_1+n-x)}{B(a_1,b_1)}\mathrm{Beta}(p \mid a_1+x,b_1+n-x)\\
&\quad+(1-\pi)\frac{B(a_2+x,b_2+n-x)}{B(a_2,b_2)}\mathrm{Beta}(p \mid a_2+x,b_2+n-x).
\end{aligned}$$

We finish by normalising the weights to obtain

$$p \mid x \sim \omega_1\mathrm{Beta}(p \mid a_1+x,b_1+n-x)+(1-\omega_1)\mathrm{Beta}(p \mid a_2+x,b_2+n-x)$$

with

$$\omega_1=\pi\frac{B(a_1+x,b_1+n-x)}{B(a_1,b_1)}\left(\pi\frac{B(a_1+x,b_1+n-x)}{B(a_1,b_1)}+(1-\pi)\frac{B(a_2+x,b_2+n-x)}{B(a_2,b_2)}\right)^{-1}$$

So the posterior is also a mixture of Beta distributions.

4.3.5 Jeffreys prior

If we do not have any prior information, it is normal to use an “uninformative” prior, i.e., a prior that assumes as little as possible about the parameter values. It is common to use uniform priors as uninformative priors, so that the posterior basically corresponds to the likelihood of the data. This is approach taken for many parameters in parameter estimation of gravitational wave data and was in fact the approach that Bayes himself advocated. However, uniform priors are not invariant under re-parameterisation. If one is ignorant about the value of θ, one is also ignorant about the value of θ^2 or any other function of θ. Therefore, any uninformative prior should induce the same form of uninformative prior on any other variables defined by transformation. Jeffreys (1961) proposed a class of priors that are invariant under re-parameterisations. By identifying the probability density with a metric on parameter space he argued that the prior should take the form $[\det(g_{ij})]^{1/2}$ where the metric

$$g_{ij}(\vec{\theta})=\frac{1}{f(\vec{\theta})}\frac{\partial f}{\partial\theta_i}\frac{\partial f}{\partial\theta_j}.$$

This would lead to an invariant prior for any scalar function $f(\vec{\theta})$. Jeffreys advocated the use of the likelihood, which introduces a data dependence into the expression, that can be eliminated by taking the expectation over realisations of the data. This procedure leads to **Jeffreys prior** which is

$$p(\vec{\theta}) \propto \sqrt{\det[I(\vec{\theta})]}, \qquad \text{where } I(\vec{\theta})_{ij}=\mathbb{E}\left[\frac{\partial l}{\partial\theta_i}\frac{\partial l}{\partial\theta_j}\right]$$

for $l=\log p(\mathbf{x}|\vec{\theta})$ the log-likelihood is the Fisher information matrix.

Jeffreys prior is “uninformative” because it can be interpreted as being as close as possible to the likelihood function and it is invariant under re-parameterisation. However, it is rarely a member of the conjugate family of distributions or of some other convenient form which is why it is not always convenient to use it in practice. Note also that the Jeffreys prior is not always **proper**, i.e., it does not always have a finite integral and therefore may not be normalisable.

Example: Poisson distribution For a single observation, x, from the Poisson(λ) distribution with pmf

$$p(x|\lambda) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}$$

we have

$$\frac{\partial \log p}{\partial \lambda} = \frac{x}{\lambda} - 1, \quad \frac{\partial^2 \log p}{\partial \lambda^2} = -\frac{x}{\lambda^2} \quad \Rightarrow \quad I(\lambda) \equiv \mathbb{E}\left[-\frac{\partial^2 \log p}{\partial \lambda^2}\right] = \frac{1}{\lambda}.$$

The Jeffreys prior for the Poisson distribution is therefore $p(\lambda) \propto 1/\sqrt{\lambda}$. This is an example of an **improper** prior, since it cannot be normalised to integrate to 1 unless the range of rates is restricted.

4.4 Posterior summary statistics

The result of a Bayesian inference calculation is a probability distribution, the full posterior probability distribution of the parameters, $p(\vec{\theta}|\mathbf{x})$. This is not only difficult to calculate in many cases, it is also unwieldy to manipulate and so it is common to use quantities that summarise the properties of the distribution. These are all of the summary statistics that we encountered in the first chapter of the course.

4.4.1 Point estimates

To obtain point estimates of a parameter value, θ_1 say, one typically works with the **marginalised** distribution for that parameter, defined by

$$p_{\mathrm{marg}}(\theta_1|\mathbf{x}) = \int p(\vec{\theta}|\mathbf{x}) \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_m.$$

From this marginal distribution, we can evaluate the **posterior mean**

$$\mu = \int_{-\infty}^{\infty} \theta_1 p_{\mathrm{marg}}(\theta_1|\mathbf{x}) \mathrm{d}\theta_1$$

or the **posterior median**, m, defined such that

$$\int_{-\infty}^{m} p_{\mathrm{marg}}(\theta_1|\mathbf{x}) \mathrm{d}\theta_1 = 0.5 = \int_{m}^{\infty} p_{\mathrm{marg}}(\theta_1|\mathbf{x}) \mathrm{d}\theta_1$$

or the **posterior mode**

$$M = \mathrm{argmax}\; p_{\mathrm{marg}}(\theta_1|\mathbf{x}).$$

The posterior mean and mode can be defined unambiguously over the full distribution as well. The posterior mean is the same whether computed over the marginal distribution or the full distribution, but the mode typically changes. The median is not unambiguously defined on the whole distribution, as there are infinitely many ways to partition the full parameter space into equal probability subsets.

4.4.2 Credible intervals

To move beyond point estimates, it is natural to want to describe ranges in which parameter values are estimated to lie. The Bayesian equivalent of a frequentist confidence interval is a **credible interval**. This is defined as

Definition: An interval (a, b) is a $100(1-\alpha)\%$ posterior credible interval for θ_1 if

$$\int_a^b p_{\mathrm{marg}}(\theta_1|\mathbf{x})\mathrm{d}\theta_1 = (1-\alpha), \quad 0 \leq \alpha \leq 1.$$

A **credible region** can be defined in a similar way. This is any partition of parameter space that contains $100(1-\alpha)\%$ of the total posterior probability. Clearly credible intervals and regions are not unique, but there are two types of credible interval that are commonly used.

Definition: An interval (a, b) is a **symmetric** $100(1-\alpha)\%$ posterior credible interval for θ_1 if

$$\int_{-\infty}^a p_{\mathrm{marg}}(\theta_1|\mathbf{x})\mathrm{d}\theta_1 = \frac{\alpha}{2} = \int_b^\infty p_{\mathrm{marg}}(\theta_1|\mathbf{x})\mathrm{d}\theta_1.$$

Definition: An interval (a, b) is a $100(1-\alpha)\%$ **highest posterior density (HPD) interval** for θ_1 if

1. $[a, b]$ is a $100(1-\alpha)\%$ credible interval for θ_1;
2. for all $\theta \in [a, b]$ and $\theta' \notin [a, b]$ we have $p_{\mathrm{marg}}(\theta|\mathbf{x}) \geq p_{\mathrm{marg}}(\theta'|\mathbf{x})$.

Credible intervals are more intuitive than confidence intervals as they make an explicit statement about the probability that the parameter takes values in the range, rather than referencing an ensemble of similar experiments.

4.4.3 Posterior samples

Summary statistics provide a useful way to summarise and compare distributions, but they inevitably discard information. To retain full information about the parameters we need the full posterior. Often this cannot be written down in a simple analytic form, but it can be summarised by drawing a set of samples $\{\vec{\theta}_1, \ldots, \vec{\theta}_M\}$ randomly from the posterior. Such samples can then be used to compute integrals over the posterior

$$\int f(\vec{\theta})p(\vec{\theta}|\mathbf{x})\mathrm{d}\vec{\theta} \approx \frac{1}{M}\sum_{i=1}^{M} f(\vec{\theta}_i).$$

Most quantities that one might wish to compute from a posterior distribution can be expressed as integrals of this form, and so generation of such samples is the most complete way to represent posterior distributions. Efficient production of samples is non-trivial and will be the topic of the next chapter of these notes.

4.5 Interpreting summary statistics

4.5.1 Decision theory

The posterior mean, mode and median are all valid ways to summarise a posterior distribution. One way to motivate these (and other possible) choices is through **decision theory**. In decision theory, understanding which decision is the best is motivated by introducing a **loss function** which characterises the cost or penalty of making a particular decision. Formally we define various quantities

- The **sample space** $\mathcal{X}$ denotes the possible values for the observed data, $\mathbf{x}$.
- The **parameter space**, Ω_θ, denotes possible (unknown) states of nature (or parameter values characterising the true pdf of observed data sets).
- We define a **family of probability distributions**, $\{\mathbb{P}_\theta(x) : x \in \mathcal{X}, \theta \in \Omega_\theta\}$, which describe how the observed data is generated in the possible states of nature.
- The **action space**, $\mathcal{A}$, is the set of actions that an experimenter can take after observing data, e.g., reject or accept a null hypothesis, assign an estimate to the value of θ etc.
- The **loss function**, $L : \Omega_\theta \times \mathcal{A} \to \mathbb{R}$, is a mapping from the space of actions and parameters to the real numbers, such that $L(a, \theta)$ is the loss associated with taking the action a when the true state of nature is θ.
- The set of **decision rules**, $\mathcal{D}$, is a set of mappings from data to actions. Each element $d \in \mathcal{D}$ is a function $d : \mathcal{X} \to \mathcal{A}$ that associates a particular action with each possible observed data set.

For a parameter value $\theta \in \Omega_\theta$, the risk of a decision rule, d, is defined as

$$R(\theta, d) = \mathbb{E}_\theta L(\theta, d(X)) = \left\{ \begin{array}{ll} \sum_{x\in\mathcal{X}} L(\theta, d(x))p(x;\theta) & \text{for discrete } X \\ \int_{\mathcal{X}} L(\theta, d(x))p(x;\theta)\mathrm{d}x & \text{for continuous } \mathcal{X}. \end{array} \right.$$

In other words, the risk is the expected loss of a particular decision rule when the true value of the unknown parameter is θ. Note that this is fundamentally a frequentist concept, since the definition implicitly invokes the idea of repeated samples from the parameter space $\mathcal{X}$ and computes the average loss over these hypothetical repetitions. However, it is possible to extend these ideas to a Bayesian framework by defining a prior, $\pi(\theta)$, over the parameters of the distribution. The **Bayes risk** of a decision rule, d, is then defined as

$$r(\pi, d) = \int_{\theta\in\Omega_\theta} R(\theta, d)\pi(\theta)\mathrm{d}\theta,$$

or by a sum in the case of a discrete-valued probability distribution. A decision rule is **a Bayes rule** with respect to the prior $\pi(\cdot)$ if it minimizes the Bayes risk, i.e.,

$$r(\pi, d) = \inf_{d'\in\mathcal{D}} r(\pi, d') = m_\pi, \text{ say.}$$

Note that, as usual in a Bayesian context, the Bayes rule depends on the specification of the prior and therefore there will be infinitely many Bayes rules for any particular problem. A useful choice of prior is the one that is most conservative in its estimate of risk. This gives rise to the concept of **a least favourable prior**. The prior $\pi(\theta)$ is least favourable if, for any other prior $\pi'(\theta)$ we have

$$r(\pi, d_\pi) \geq r(\pi', d_{\pi'})$$

where d_π, $d_{\pi'}$ are the Bayes rules corresponding to $\pi(\cdot)$ and $\pi'(\cdot)$ respectively.

4.5.2 Bayes rules as minimizers of posterior expected loss

The Bayes risk can be written as

$$\begin{aligned} r(\pi, d) &= \int_{\Omega_\theta} R(\theta, d)\pi(\theta)\mathrm{d}\theta \\ &= \int_{\Omega_\theta}\int_{\mathcal{X}} L(\theta, d(x))p(x|\theta)\pi(\theta)\mathrm{d}x\mathrm{d}\theta \\ &= \int_{\Omega_\theta}\int_{\mathcal{X}} L(\theta, d(x))p(\theta|x)p(x)\mathrm{d}x\mathrm{d}\theta \\ &= \int_{\mathcal{X}} p(x)\left\{\int_{\Omega_\theta} L(\theta, d(x))p(\theta|x)\mathrm{d}\theta\right\}\mathrm{d}x \end{aligned}$$

where the second line follows from the definition of the risk function and the third line follows by using Bayes' theorem to write $p(x|\theta)\pi(\theta) = p(\theta|x)p(x)$ in terms of the posterior $p(\theta|x)$ and the evidence $p(x)$. The Bayes rule minimizes the Bayes risk. We see that this minimum is achieved for a particular value of x by making the decision that minimizes the expression in curly brackets. This is the **expected posterior loss** associated with the observed x. This observation simplifies the calculation in many cases and also illustrates the general property of Bayesian procedures, namely that the decision depends only on the observed data and not on potential unobserved data sets.

We will illustrate this with four examples. In the first three examples, we are attempting to make a point estimate and so the decision is an assignment of the value of the parameter $d = \hat{\theta}$.

Example: Point estimation with squared error loss Suppose we want to make a point estimate of a parameter and we use a squared error loss function, $L(\theta, d) = (\theta - d)^2$. Find the Bayes rule.

Solution

The Bayes rule chooses $d(Y)$ to minimize

$$\int_{\Omega_\theta} (\theta - d)^2 p(\theta|y)\mathrm{d}\theta.$$

Differentiating with respect to d and setting this to zero gives

$$\int_{\Omega_\theta} (\theta - d)p(\theta|x)\mathrm{d}\theta = 0 \quad \Rightarrow \quad d = \int_{\Omega_\theta} \theta p(\theta|x)\mathrm{d}\theta.$$

In other words, the Bayes estimator of θ, with squared error loss, is the **posterior mean**.

Example: Point estimation with absolute magnitude error loss
Suppose we instead use the loss function $L(\theta, d) = |\theta - d|$. Find the new Bayes rule.

Solution
In this case, the Bayes rule minimizes

$$\int_{-\infty}^{d} (d-\theta) p(\theta|x) \mathrm{d}\theta + \int_{d}^{\infty} (\theta - d) p(\theta|x) \mathrm{d}\theta.$$

Setting the derivative with respect to d to zero now gives

$$\int_{-\infty}^{d} p(\theta|x) \mathrm{d}\theta - \int_{d}^{\infty} p(\theta|x) \mathrm{d}\theta = 0 \quad \Rightarrow \quad \int_{-\infty}^{d} p(\theta|x) \mathrm{d}\theta = \int_{d}^{\infty} p(\theta|x) \mathrm{d}\theta = \frac{1}{2}.$$

In other words, the Bayes estimator of θ, with absolute magnitude error loss, is the **posterior median**.

Example: Point estimation with delta-function gain
Suppose we instead use the loss function

$$L(\theta, d) = \begin{cases} -\delta(\theta - d) & \text{if } d = \theta \\ 0 & \text{if } d \neq \theta \end{cases}.$$

In other words, the loss is infinitely higher for any value except the correct one. Find the new Bayes rule.

Solution
In this case, the Bayes rule minimizes

$$-\int_{-\infty}^{\infty} \delta(\theta - d) p(\theta|x) \mathrm{d}\theta = -p(d|x).$$

The minimum loss is obtained by setting

$$d = \mathrm{argmaxp(d|x)},$$

i.e., the posterior mode.

Example: Interval estimation
Suppose we have a loss function of the form

$$L(\theta, d) = \begin{cases} 0 & \text{if } |\theta - d| \leq \delta \\ 1 & \text{if } |\theta - d| > \delta \end{cases}$$

for specified $\delta > 0$. What is the Bayes rule?

Solution

The expected posterior loss in this case is the posterior probability that $|\theta - d| > \delta$. The interval that minimises this loss, among intervals of fixed length 2δ, is the interval that contains the highest posterior probability. This is called the *highest posterior density* interval.

We see that all of the "natural" ways to obtain a point estimate from a Bayesian posterior can be interpreted in terms of Bayes rule's with different loss functions.

4.6 Bayesian hypothesis testing

The denominator that appears in Bayes' theorem is the Bayesian evidence and can be computed via

$$\mathcal{Z} = p(\mathbf{x}) = \int p(\mathbf{x} \mid \vec{\theta})p(\vec{\theta})\mathrm{d}\vec{\theta}.$$

When writing down Bayes' theorem we suppressed the fact that all of the quantities were conditioned on the particular model we were assuming for the data generating process. Explicitly reintroducing the dependence on the model, M, we have

$$p(\vec{\theta}|\mathbf{x}, M) = \frac{p(\mathbf{x}|\vec{\theta}, M)p(\vec{\theta}|M)}{p(\mathbf{x}|\mathbf{M})}.$$

This makes it clear that the evidence, $p(\mathbf{x}|\mathbf{M})$, represents the *probability of seeing the model data under model M* and can be thought of as the likelihood for the model given the observed data. If we now have more than one model, M_1 and M_2 say, that we believe could describe the data, we can compute the **posterior odds ratio** for M_1 over M_2

$$O_{12} = \frac{p(\mathbf{x}|\mathbf{M_1})}{p(\mathbf{x}|\mathbf{M_2})}\frac{p(M_1)}{p(M_2)}.$$

The first term is called the **Bayes factor** and is the ratio of the model likelihoods. The second term is the **prior odds ratio**, which represents our prior belief about the relative probability of the two models. The posterior odds is the ratio of model probabilities based on the observed data and is the basis for Bayesian hypothesis testing. For $O_{12} \gg 1$ we favour model M_1, while for $O_{12} \ll 1$ we favour M_2.

In the case of a flat prior on models the prior odds ratio is just 1 and decisions are based on the Bayes factor. Kass and Rafferty (1995) described a 'rule of thumb' for interpreting Bayes' factors. This is summarised in Table 2. This Table can be used to interpret the results of Bayesian hypothesis tests. Alternatively, the distribution of the Bayes factor can be computed under the null hypothesis and used, in a frequentist way, to produce a mapping between p-values and Bayesian posterior odds ratios.

The models M_1 and M_2 need not be very different, but could, for example, represent different regions of the parameter space of a distribution, e.g., $M_1 : \theta \in \Theta_1$ versus $M_2 : \theta \in \Theta_2$. If the two hypotheses are both simple then the Bayes factor reduces to the likelihood ratio, which we saw was the optimal test statistic in the frequentist hypothesis testing context.

Computation of the Bayesian evidence is challenging. Most sampling algorithms that return independent samples from the posterior ignore the evidence as it is just a normalisation constant. The evidence can be written as an integral over the posterior which can be

Bayes Factor	Interpretation
< 3	No evidence of M_1 over M_2
> 3	Positive evidence for M_1
> 20	Strong evidence for M_1
> 150	Very strong evidence for M_1

Table 2: Table for intepretation of Bayes' factors, as presented in Kass and Rafferty (1995).

approximated by a sum over samples

$$\frac{1}{\mathcal{Z}} = \int \frac{1}{p(\mathbf{x} \mid \vec{\theta})} \frac{p(\mathbf{x} \mid \vec{\theta})p(\vec{\theta})}{\mathcal{Z}} \mathrm{d}\vec{\theta} \approx \frac{1}{M} \sum_{i=1}^{M} \frac{1}{p(\mathbf{x} \mid \vec{\theta}_i)}.$$

In other words it is the harmonic mean of the likelihoods of the samples. This is an extremely unstable approximation, however, as this sum is dominated by points with small likelihoods, but these are precisely the regions where there will be fewer samples and hence larger Monte Carlo error. Other techniques, such as nested sampling, can be used to compute evidences more accurately and these will be discussed in the next chapter.

Example: Suppose we have a two dimensional Normal likelihood of the form

$$p(\mathbf{x}|\vec{\theta}) = \frac{\sqrt{1-\rho^2}}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{1}{} \left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + 2\frac{\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right] \tag{67}$$

and use priors for the parameters μ_1 and μ_2 of the form

$$p(\mu_1) = \frac{1}{\Sigma_1\sqrt{2\pi}} \exp\left[-\frac{1}{2\Sigma_1^2}\mu_1^2\right], \qquad p(\mu_2) = \frac{1}{\Sigma_2\sqrt{2\pi}} \exp\left[-\frac{1}{2\Sigma_2^2}\mu_2^2\right]. \tag{68}$$

We are interested in comparing the two models

$$M_1 : \mu_2 = 0, \qquad M_2 : \mu_2 \in (-\infty, \infty).$$

The evidence for M_1 can be computed as

$$\mathcal{Z}_1 = \frac{1}{2\pi\sigma_2}\sqrt{\frac{1-\rho^2}{\sigma_1^2+\Sigma_1^2}} \exp\left[-\frac{x_2^2(\sigma_1^2-(1-\rho^2)\Sigma_1^2) + 2\rho x_1 x_2\sigma_1\sigma_2 + \sigma_2^2 x_1^2}{2\sigma_2^2(\sigma_1^2+\Sigma_1^2)}\right]$$

and for M_2 it is

$$\begin{aligned}\mathcal{Z}_2 = \frac{1}{2\pi}&\sqrt{\frac{1-\rho^2}{\sigma_1^2(\sigma_2^2+\Sigma_2^2)+\Sigma_1^2(\sigma_2^2+(1-\rho^2)\Sigma_2^2)}}\times \\ &\times \exp\left[-\frac{x_2^2((1-\rho^2)\Sigma_1^2+\sigma_1^2) + 2\rho x_1 x_2\sigma_1\sigma_2 + x_1^2((1-\rho^2)\Sigma_2^2+\sigma_2^2)}{2\Sigma_1^2((1-\rho^2)\Sigma_2^2+\sigma_2^2) + 2\sigma_1^2(\sigma_2^2+\Sigma_2^2)}\right]\end{aligned} \tag{69}$$

which gives the posterior odds ratio in favour of M_2, for equal prior odds (which is just the Bayes factor)

$$\begin{aligned}\mathcal{O}_{21} = \frac{\mathcal{Z}_2}{\mathcal{Z}_1} = \sigma_2&\sqrt{\frac{\Sigma_1^2+\sigma_1^2}{\Sigma_1^2((1-\rho^2)\Sigma_2^2+\sigma_2^2)+\sigma_1^2(\Sigma_2^2+\sigma_2^2)}}\times \\ &\times \exp\left[\frac{\Sigma_2^2(x_2((1-\rho^2)\Sigma_1^2+\sigma_1^2)+\rho x_1\sigma_1\sigma_2)^2}{2(\Sigma_1^2+\sigma_1^2)\sigma_2^2(\sigma_1^2(\Sigma_2^2+\sigma_2^2)+\Sigma_1^2((1-\rho^2)\Sigma_2^2+\sigma_2^2))}\right].\end{aligned} \tag{70}$$

This is difficult to interpret, but if we now assume that $\Sigma_1^2 \gg \sigma_1^2$, i.e., that the prior in μ_1 is much broader than the typical measurement uncertainty, the odds ratio simplifies to

$$\mathcal{O}_{21} \approx \sigma_2 \sqrt{\frac{1}{(1-\rho^2)\Sigma_2^2 + \sigma_2^2}} \exp\left[\frac{(1-\rho^2)x_2^2}{2\sigma_2^2}\right]$$

We see that there is a competition between the size of the additional variable dimension (characterised by Σ_2) in the first term and the weight of evidence for the additional effect in the data (characterised by the second term). Only if the addition of the extra dimension significantly improves the fit to the data (characterised by x_2 which is effectively the peak of the posterior in μ_2 when that parameter is allowed to vary) should the more complex model be favoured. If the fit does not improve, then the addition of the extra dimension is penalised by the first term and so the more complex model should not be preferred. It is often said that Bayesian posterior odds ratios automatically encode the notion of "Occam's razor", i.e., one should use the simplest model that adequately describes the data since adding extra degrees of freedom always improves a fit. This is the sense in which it is meant. Addition of extra dimensions typically includes a prior penalty, as we see here, which will lead to the disfavouring of an alternative model unless the likelihood shows a significantly great improvement when the extra degrees of freedom are included.

4.7 Predictive checking

In both a frequentist and a Bayesian context it is natural to ask whether the model is a good representation of the observed data. In the Bayesian context this is accomplished by using **predictive distributions**.

Definition: the **prior predictive distribution** is the probability distribution

$$p(\mathbf{x}) = \int_{\vec{\theta}\in\Theta} p(\mathbf{x}|\vec{\theta})p(\vec{\theta})\mathrm{d}\vec{\theta}.$$

This is the likelihood weighted by the assigned prior distribution and therefore represents our *a priori* belief about the distribution of data sets that would be observed. Similarly, we have the following

Definition: the **posterior predictive distribution** is the probability distribution

$$p(\mathbf{y}|\mathbf{x}) = \int_{\vec{\theta}\in\Theta} p(\mathbf{y}|\vec{\theta})p(\vec{\theta}|\mathbf{x})\mathrm{d}\vec{\theta}.$$

This is the likelihood weighted by the posterior probability based on the observed data $\mathbf{x}$ and is our expectation about the distribution of future data sets $\mathbf{y}$.

The posterior predictive distribution can be used to assess whether the observed data is unusual within the posterior distribution, which is an indicator about whether or not the model is a good fit. Based on the observed data $\mathbf{x}$ we generate a large number of new data sets $\{\mathbf{y}_1, \ldots \mathbf{y}_N\}$ that are similar to $\mathbf{x}$, i.e., they consist of the same number of observations. For each data set we compute a set of summary statistics, and hence obtain the distribution of the summary statistics over many realisations of the posterior predictive distribution. We can then assess the "p-value" of the observed data within these distributions. If it looks like an outlier in any one of these distributions this suggests the model is not a good fit. Suitable summary statistics could include the maximum, minimum, median, skewness, kurtosis etc. Ideally we choose summary statistics that are orthogonal to the model parameters to increase sensitivity, since we are using the data twice (once to compute the posterior and once to compare to the predictive distribution). Statistics that are effectively tuned to the observed data will tend to lie in the middle of the predictive distributions by construction, even if the model is poor. We will see an example of this in the next section.

4.8 Example: regression

To illustrate some of the ideas discussed above we will present a Bayesian analysis of a regression problem. We suppose that we have made measurements of a set of values, $\{y_i\}$, corresponding to sets of p known explanatory variables, $\{\mathbf{x}_i\}$, and we believe that these follow a linear relationship with equal variance normally distributed errors

$$y_i \sim N(\mathbf{x}_i^T \vec{\beta}, \sigma^2), \quad i = 1, \ldots, N.$$

We want to infer the parameters of the linear relationship, $\vec{\beta}$, and the unknown precision $\tau = 1/\sigma^2$. We use a Bayesian framework and so must write down prior distributions on these parameters. We can assume a separable prior

$$p(\vec{\beta}, \tau) = p(\tau) \prod_{i=1}^{p} p(\beta_j)$$

and take Normal priors for the β_j's and a Gamma prior for τ as these are conjugate priors in the Normal-Gamma model

$$\beta_j \sim N(\mu_{\beta_j}, \sigma^2_{\beta_j}), \quad \tau \sim \text{Gamma}(a, b).$$

In the absence of prior information it is reasonable to set $\mu_{\beta_j} = 0$. Inferred values of the coefficients that are non-zero then provide evidence for the existence of a relationship between the observed data and those explanatory variables. Setting σ_j^2 to a large value, say 10^4, indicates large uncertainty in the parameter values and avoids strong prior dependence in the results. For the prior on τ, it is usual to take small values of a and b, for example $a = b = 0.1$ or $a = b = 0.01$. However, such priors lead to a preferred value (i.e., a peak) in the prior and so the use of such priors is somewhat controversial.

To illustrate fitting such a model, we can use a standard data set, the MTCARS data set, which is available in the R statistical software package and may also be found online. The data set contains observations, y_i, of the miles driven per gallon in the i'th of 32 different models of car, with explanatory variables x_{i1}, the rear axle ratio, x_{i2}, the weight of the i'th car and x_{i3}, the time to drive 0.25 miles from rest. We fit the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, 1/\tau), \quad i = 1, \ldots 32,$$

with $\beta_j \sim N(0, 1000)$ and $\tau \sim \text{Gamma}(0.1, 0.1)$. We can use statistical software (in this case R) to generate samples from the posterior. Techniques for doing this will be discussed in the next chapter, and in the associated practical. using these samples we can obtain a posterior mean and 95% symmetric credible interval for each parameter. These can be compared to the frequentist estimates of the same parameters and the frequentist 95% confidence interval (see problem sheet 1). This comparison is in Table 3.

	Bayesian results		Frequentist results	
Parameter	Posterior mean	95% credible interval	MLE	95% confidence interval
β_0	10.369	[-5.098,36.349]	11.395	[-5.134,27.922]
β_1	1.777	[-0.721,4.166]	1.750	[-0.857,4.169]
β_2	-4.335	[-5.702,-2.995]	-4.347	[-5.787,-3.009]
β_3	0.968	[0.449,1.493]	0.946	[0.410,1.482]
σ^2	6.978	[4.160,11.729]	6.554	—

Table 3: Comparison between Bayesian and frequentist estimates of the linear model fit to the MTCARS data set.

The results of the Bayesian fit are quite consistent between the two approaches, although there are some differences and the interpretation of the results is different. We now want to assess the quality of the results. In a frequentist setting, assessment of the quality of a linear model fit is done through the production of *studentised residuals* and *Q-Q plots.* A studentised residual is

$$\hat{\epsilon}_i = \frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

where $\hat{\beta}$ are the estimated parameters, $\hat{\sigma}$ is the esitmated standard deviaiton and h_{ii} is the i'th diagonal element of the matrix $H = \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T$. These quantities follow a student-t distribution which is why they are called studentised residuals. A $Q - Q$ plot is a plot of the distribution of these values against the theoretical distribution, which should be approximately a straight line if the model is a good description of the data.

We can construct analogous quantities in the Bayesian case, but now the parameters are described by distributions rather than point estimates. A point estimate can be constructed in a number of different ways — using posterior mean values, using a single draw from the posterior, or averaging over the full posterior. The latter approach involves computing the studentised residual for a large number of draws from the posterior and averaging them, and is called the *posterior mean of the residual.* Studentised residuals are plotted in various ways in Figure 5.

We can also produce posterior predictive checks as described in section 4.7. We compute realisations of similar data sets and estimate the distribution of various summary statistics which we then compare to the values in the observed data sets. In this case we compute the distributions of the minimum, maximum, median and skewness in repeated data sets. These are shown in Figure 6, along with the values in the observed data set. We see that the observed values lie within the distributions in all cases, except for skewness. Seeing that the observed data lies in the tail of the distribution may indicate a failure of the model. In this case we might want to try varying the assumption of normally distributed errors and homoskedacity (equal error variance).

The issue with the posterior predictive checks could indicate a failure of the model, or the influence of an outlying data point. One way to tackle this is to modify the model so that the distribution of the errors ϵ_i is no longer assumed to be normal. The most common approach is to replace the normal distribution by a t_ν-distribution, as these have heavier tails. This is referred to as **robust regression**. The degrees of freedom, ν, in the t_ν-distribution can be fixed to some reasonable value, or allowed to vary in a hierarchical model (see next section). In that case the prior on ν is usually taken to be a Gamma distribution, $\nu \sim \text{Gamma}(c, d)$.

For the MTCARS dataset we try this, using prior values $c = d = 0.1$, and then look at the

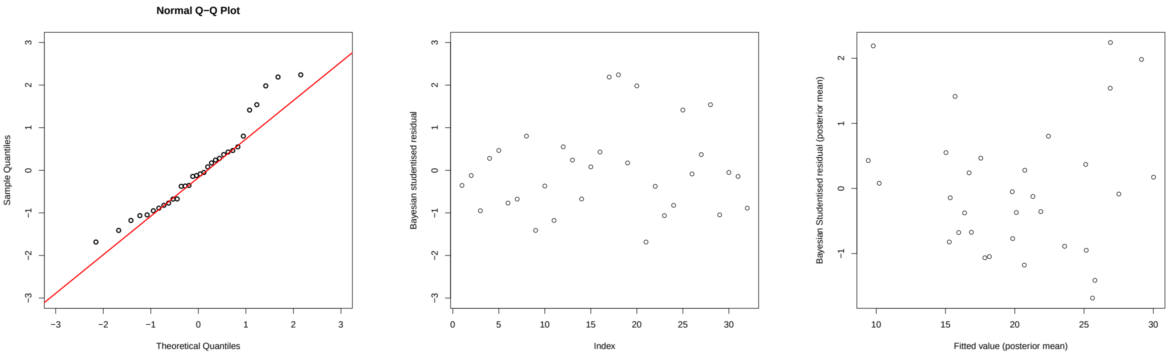


Figure 5: $Q-Q$ plot of the studentised residuals (left), studentised residual versus index of data point (middle) and studentised residual versus posterior mean of the predicted value, $\hat{y}_i$, for the Bayesian fit to the MTCARS data set. We look for the left hand plot to be on the diagonal line, for the middle and right hand plots we want the values to be randomly distributed (i.e., no trend with the x value) and in the range from minus a few to plus a few. These constraints are all satisfied here and so we see no cause for concern.

posterior predictive distribution again. The results for the skewness are shown in Figure 7. We that robustifying regression can help to improve the model fit in this case. The observed dat moves from lying at the 99.6% point of the distribution to lying at the 96.3%. So, it is still something of an outlier but it is not so much a cause for concern. It is perhaps not surprising that the use of robust regression only helped a small amount in this case, since we are trying to compensate for non-zero skew in the data and the t-distribution is also a symmetric distribution.

4.9 Hierarchical models

In many contexts, for example the observation of mergers of compact binary coalescences through gravitational wave observations, the likelihood describes the observation of a single event, and the prior describes the distribution of parameter values in the population from which the events are drawn. Often the parameters of the population prior are not themselves known but are of interest. For example, we do not know the distribution of masses of black holes in binaries and would like to learn about this from observations of the gravitational wave sources. This leads to the notion of a **hierarchical model**, in which the likelihood for data depends on parameters for which we write down a prior that in turn depends on unknown parameters (usually termed **hyperparameters**), for which we write down another prior (the **hyperprior**).

This hierarchy can be continued to more and more levels, but such models increase rapidly in complexity. Inference on complex hierarchical models can be simplified by imposing a *conditional independence* structure in the models, e.g., $p(x, y, z) = p(x|z)p(y|z)p(z)$. Conditional dependence structures can be compactly represented using *graphical models*. These are directed acyclic graphs that indicate dependencies between various components of the model. It is important that the graph has no cycles as only then can the joint probability be factorised. An example of a graphical model is shown in Figure 8. This model represents the following conditional dependence structure

$$p(p, q, r, s, t, u, v, w, x, y, z) = p(x|y, z)p(y|u, w)p(w|v)p(u)p(v)p(z|r)p(r|p, q)p(p)p(q) \quad (71)$$

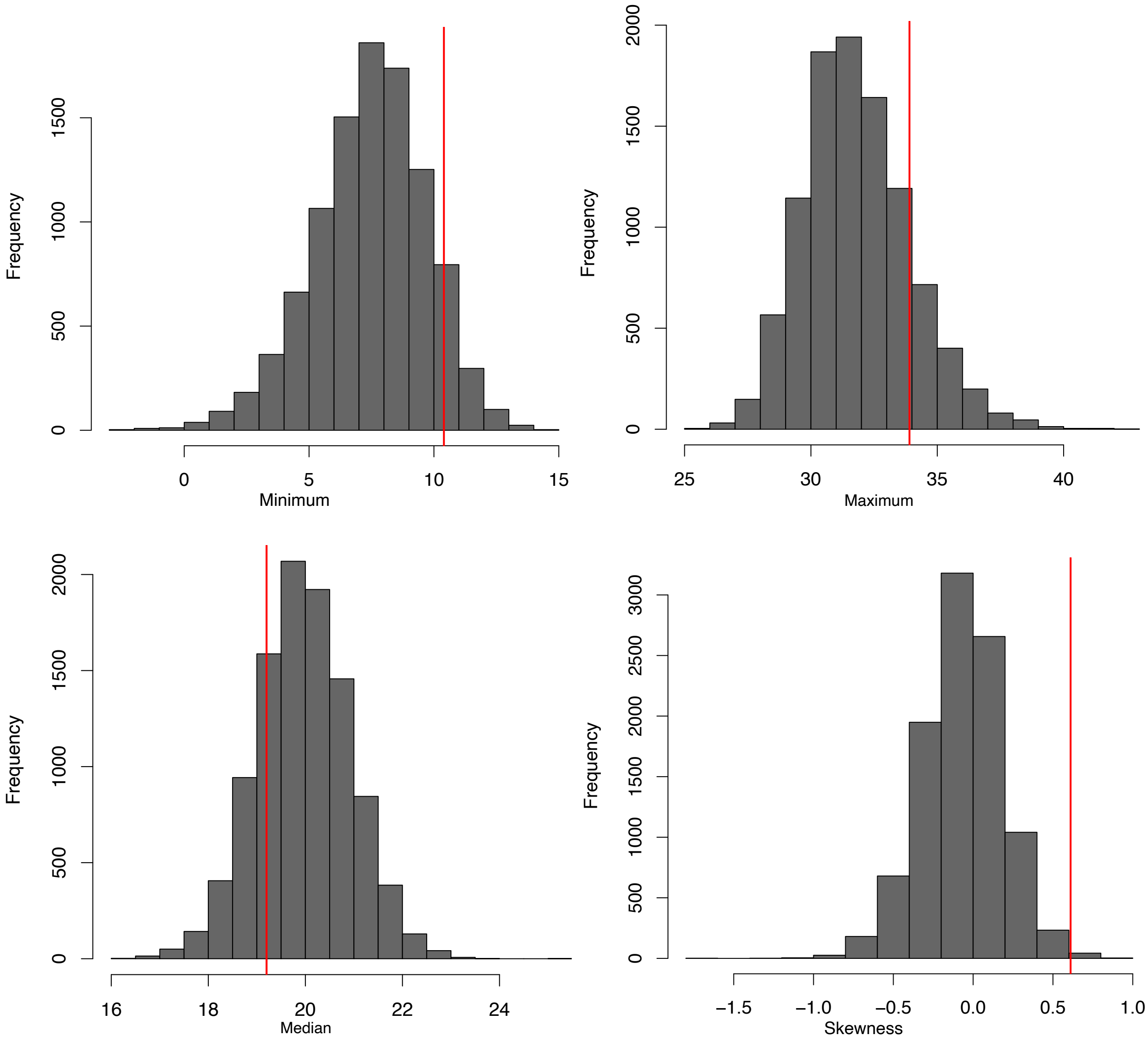


Figure 6: Predictive distributions for the maximum (top left), minimum (top right), median (bottom right) and skewness (bottom right) in replicated data sets of size 32, based on the posterior distribution from the MTCARS data set. The vertical red lines indicate the values in the data set form which the posterior was obtained. We see that this lies in the middle of the distribution in all cases, except skewness, in which it lies in the tail, which might indicate a failure to properly fit the data.

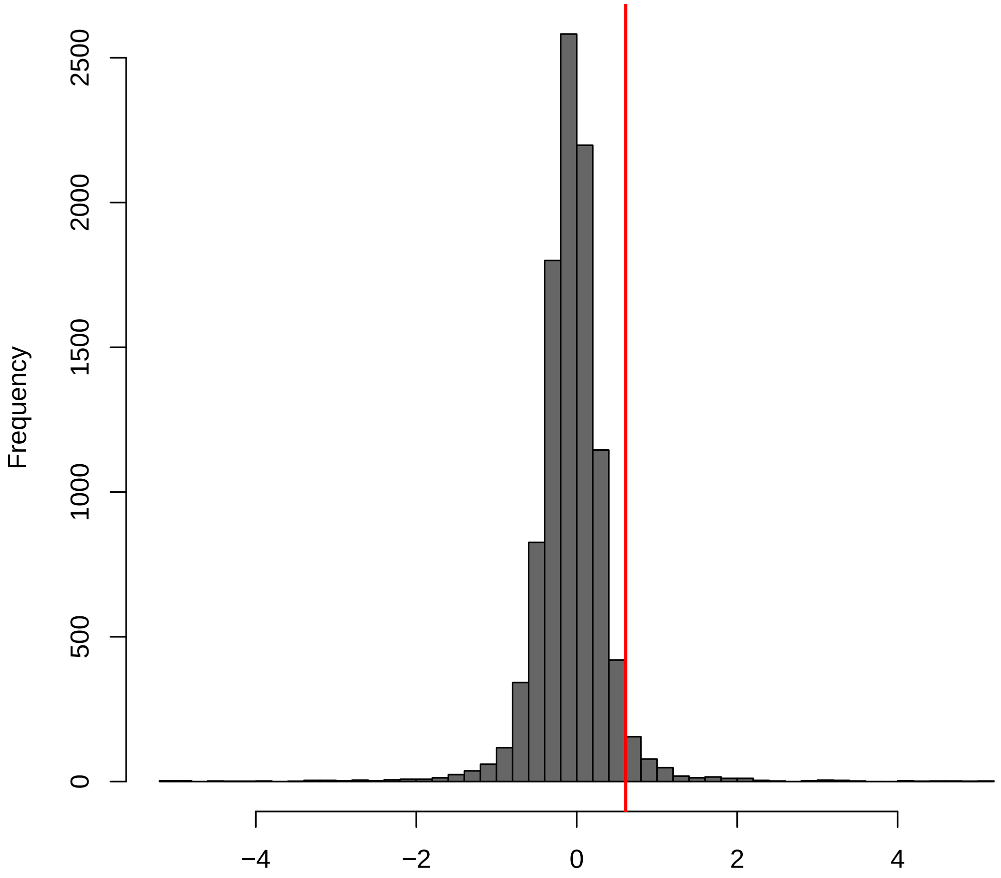


Figure 7: Posterior predictive distribution of skewness for the robustified regression model. The observed value of the skewness is indicated by a vertical red line as before.

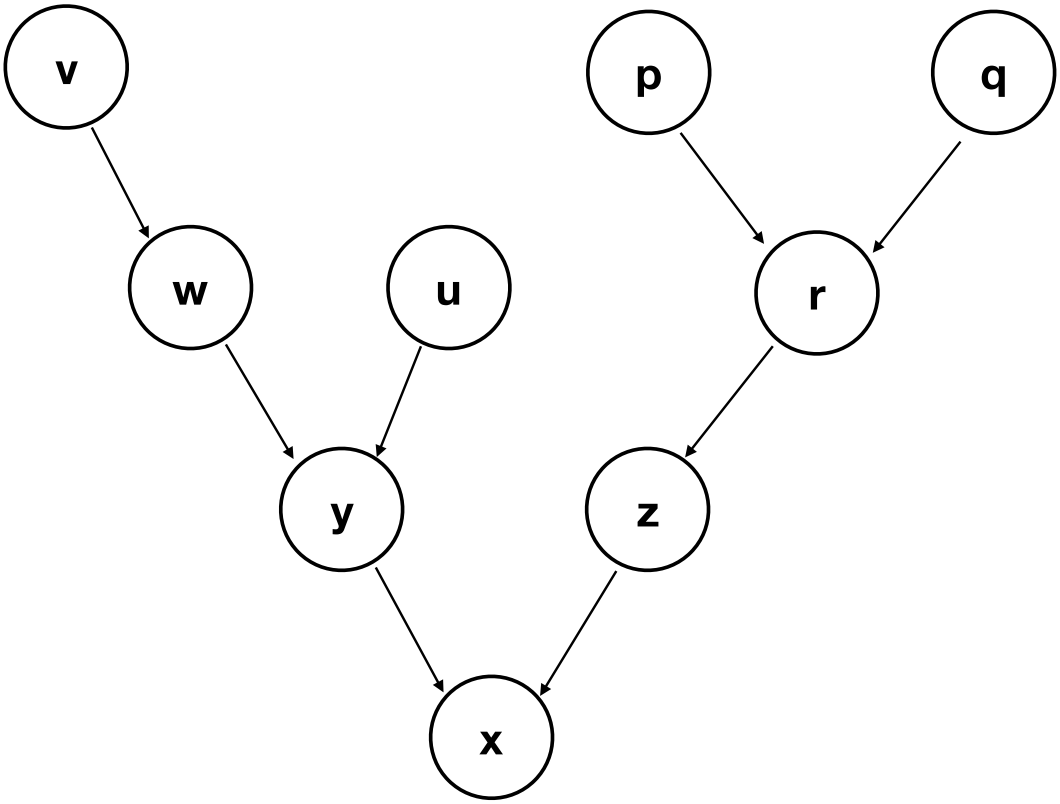


Figure 8: Illustration of a Bayesian graphical model. This is an acyclic directed graph that indicates conditional dependencies in complex Bayesian hierarchical models.

4.9.1 Selection effects

One thing that is important to account for in hierarchical modelling are selection effects. The decision about whether or not to include an event in a catalogue used for inference is based on whether or not the event is "detected", i.e., whether or not the observed data passes some pre-determined threshold criterion for inclusion. This is usually a property of the data only. Selection effects can be included by modifying the likelihood so that it represents the likelihood of "detected" data sets. If the un-corrected likelihood is $p(\mathbf{x}|\vec{\theta})$ then the likelihood for observed events is just

$$p(\mathbf{x}|\vec{\theta},\mathrm{obs}) = \frac{1}{p_s(\vec{\theta})} p(\mathbf{x}|\vec{\theta}), \quad \text{where } p_s(\vec{\theta}) = \int_{\mathbf{x}>\mathrm{threshold}} p(\mathbf{x}|\vec{\theta}) \mathrm{d}\mathbf{x}.$$

The integral is over all data sets that would have been considered as "detections", i.e., passing the threshold for inclusion in inference. What we have done here is renormalise the likelihood so that it integrates to 1 over all above threshold data sets. Since the partition of the data into observed and unobserved is a property of $\mathbf{x}$ only, the relative probabilities of different above threshold data sets must be in proportion to their probabilities in the set of all data sets.

Usually, the likelihood will depend on parameters of the particular source, $\vec{\theta}$, that are themselves determined by the priors, which depends on the hyperparameters of the population, $\vec{\lambda}$. Then the likelihood for observed events, marginalised over the source parameters is simply

$$p(\mathbf{x}|\vec{\lambda}, \mathrm{obs}) = \frac{1}{p_s(\vec{\lambda})} \int p(\mathbf{x}|\vec{\theta}) p(\vec{\theta}|\vec{\lambda}) \mathrm{d}\vec{\theta}, \quad \text{where } p_s(\vec{\lambda}) = \int_{\mathbf{x}>\text{threshold}} \int p(\mathbf{x}|\vec{\theta}) p(\vec{\theta}|\vec{\lambda}) \mathrm{d}\vec{\theta} \mathrm{d}\mathbf{x}.$$

Usually we are interested in the parameters of individual sources as well as the overall population parameters. The joint likelihood of $\mathbf{x}$ and $\vec{\theta}$, conditioned on detection, is

$$p(\mathbf{x}, \vec{\theta}|\vec{\lambda}, \mathrm{obs}) = p(\mathbf{x}|\vec{\theta}, \mathrm{obs}) p(\vec{\theta}|\vec{\lambda}, \mathrm{obs}).$$

The first term is Eq. (4.9.1), but for the source parameters $\vec{\theta}$

$$p(\mathbf{x}|\vec{\theta}, \mathrm{obs}) = \frac{p(\mathbf{x}|\vec{\theta})}{p(\mathrm{obs}|\vec{\theta})}, \quad \text{where } p(\mathrm{obs}|\vec{\theta}) = \int_{\mathbf{x}>\text{threshold}} p(\mathbf{x}|\vec{\theta}) \mathrm{d}\mathbf{x}.$$

The second term is the prior on $\vec{\theta}$ *for events above threshold.* However, this prior is modified from $p(\vec{\theta}|\vec{\lambda})$ by the conditioning on detection, namely

$$p(\vec{\theta}|\vec{\lambda}, \mathrm{obs}) = \frac{p(\vec{\theta}, \mathrm{obs}|\vec{\lambda})}{p(\mathrm{obs}|\vec{\lambda})} = \frac{p(\mathrm{obs}|\vec{\theta}, \vec{\lambda}) p(\vec{\theta}|\vec{\lambda})}{p(\mathrm{obs}|\vec{\lambda})} = \frac{p(\mathrm{obs}|\vec{\theta}) p(\vec{\theta}|\vec{\lambda})}{p_s(\vec{\lambda})}.$$

Putting this together we see that the terms relating to selection on $\vec{\theta}$, $p(\mathrm{obs}|\vec{\theta})$, cancel and the joint likelihood is

$$p(\mathbf{x}, \vec{\theta}|\vec{\lambda}, \mathrm{obs}) = \frac{p(\mathbf{x}|\vec{\theta}) p(\vec{\theta}|\vec{\lambda})}{p_s(\vec{\lambda})}$$

giving a posterior on $\vec{\theta}$

$$p(\vec{\theta}|\mathbf{x}, \vec{\lambda}, \mathrm{obs}) \propto p(\mathbf{x}|\vec{\theta}) p(\vec{\theta}|\vec{\lambda})$$

which is unchanged from the posterior that would be written down if there is no selection. We see that the selection effects corrections do not change inference about the parameters of individual sources, only inference about the hyperparameters governing the population as a whole.

This approach implicitly assumes that the number of observed events contains no information about the unknown parameters. An alternative approach is to write down a joint likelihood for all events, both the N_{obs} events that are observed, $\{\mathbf{x}_i\}$, with parameters $\{\vec{\theta}_i\}$, and the N_{nobs} events that are unobserved, $\{\mathbf{x}_j\}$, with parameters $\{\vec{\theta}_j\}$. We model the number of events as a Poisson process with overall rate $N(\vec{\lambda})$, and rate density $\mathrm{d}N/\mathrm{d}\vec{\theta}$. The joint likelihood is

$$p\left(\left\{\vec{\theta}_i\right\}, \left\{\vec{\theta}_j\right\}, \{\mathbf{x}_i\}, \{\mathbf{x}_j\} \mid \vec{\lambda}\right) \propto \left[\prod_{i=1}^{N\mathrm{obs}} p\left(\mathbf{x}_i \mid \vec{\theta}_i\right) \frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}_i}\left(\vec{\lambda}\right)\right] \times$$
$$\times \left[\prod_{j=1}^{N_{\mathrm{nobs}}} p\left(\mathbf{x}_j \mid \vec{\theta}_j\right) \frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}_j}\left(\vec{\lambda}\right)\right] \exp\left[-N\left(\vec{\lambda}\right)\right] \qquad (72)$$

We can marginalise over the unobserved data to obtain

$$p\left(\left\{\vec{\theta}_i\right\},\{\mathbf{x}_i\}\mid\vec{\lambda}\right)\propto\quad\left[\prod_{i=1}^{N_{\rm obs}}p\left(\mathbf{x}_i\mid\vec{\theta}_i\right)\frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}_i}\left(\vec{\lambda}\right)\right]\frac{N_{\rm ndet}^{N_{\rm nobs}}\left(\vec{\lambda}\right)}{N_{\rm nobs}!}\exp\left[-N\left(\vec{\lambda}\right)\right] \tag{73}$$

where

$$N_{\rm ndet}\left(\vec{\lambda}\right)\equiv\int_{\{\mathbf{x}<\text{threshold}\}}\mathrm{d}\mathbf{x}\,\mathrm{d}\vec{\theta}\,p\left(\mathbf{x}\mid\vec{\theta}\right)\frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}}\left(\vec{\lambda}\right). \tag{74}$$

We can then marginalise over the unknown number of unobserved events to obtain

$$p\left(\left\{\vec{\theta}_i\right\},\{\mathbf{x}_i\}\mid\vec{\lambda}\right)\propto\left[\prod_{i=1}^{N_{\rm obs}}p\left(\mathbf{x}_i\mid\vec{\theta}_i\right)\frac{\mathrm{d}N}{\mathrm{d}\vec{\theta}_i}\left(\vec{\lambda}\right)\right]\exp\left[-N_{\rm det}\left(\vec{\lambda}\right)\right]. \tag{75}$$

We can now introduce the overall rate in the Unvierse, N, by writing $\mathrm{d}N/\mathrm{d}\vec{\theta}=Np(\vec{\theta}|\vec{\lambda})$. Then

$$N_{\rm det}(\vec{\lambda})=N\int_{\mathbf{x}>\text{threshold}}\int p(\mathbf{x}|\vec{\theta})p(\vec{\theta}|\vec{\lambda})\mathrm{d}\vec{\theta}\mathrm{d}\mathbf{x}=Np_s(\vec{\lambda}). \tag{76}$$

Setting a scale-invariant prior on N (which states that the number of detected events does not convey information about the unknown parameters of the population), $p(N)\propto 1/N$ we can marginalise N out of the likelihood and recover Eq. (4.9.1).

4.9.2 Examples of hierarchical models

We finish this section with two examples of Bayesian hierarchical models.

Example 1: Salmon fishery In a given year, several fish hatcheries located along rivers in Washington state, USA raise coho salmon from eggs to a juvenile stage. Each hatchery releases a batch of juvenile fish into the rivers. The fish then travel to the ocean and some of them return to the hatchery 3 years later. The probability that a juvenile salmon returns varies between hatcheries due to different hatchery practices and river conditions at the point of release. We construct a hierarchical model for this as follows

- Suppose there are J fisheries and n_j salmon observed at fishery j.
- The data for an individual observation, x_{ji}, of the i'th salmon at fishery j is Bernoulli (salmon returned or did not return), with parameter p_j, where j labels the fishery. The data for the total number of returning salmon at site j, x_j, is Binomial with parameters (n_j,p_j).
- We assume that the p_j's are drawn from some common global distribution and use the conjugate prior of Beta(a,b).
- The parameters a and b are not known and fixed as in the usual case, but these are unknown quantities of interest as they characterise the variability in the population. These are the hyperparameters of the prior on p_j.
- We define a suitable hyperprior $p(a,b)$ on the hyperparameters, for example a Gamma prior.

- The joint posterior on the set $(\{p_j\}, a, b)$ is

  $$p(\{p_j\}, a, b|\mathbf{x}) \propto p(\mathbf{x}|\{p_j\}) \left[\prod_{j=1}^{J} p(p_j|a, b)\right] p(a, b).$$

  Note that the hyperprior on the hyperparameters appears only once as these parameters are common to all of the individual observations of fisheries.

- The marginal distribution on the hyperparameters (a, b) can be found by marginalising over the $\{p_j\}$'s

  $$p(a, b|\mathbf{x}) \propto p(a, b) \prod_{j=1}^{J} \frac{B\left(a + x_j, b + n_j - x_j\right)}{B(a, b)}.$$

- Marginals on individual p_j's can be found in a similar way.

Example 2: Gravitational wave cosmology In August 2017 the LIGO/Virgo gravitational wave detectors observed gravitational waves from the inspiral and merger of a binary neutron star for the first time, GW170817. There was both a short gamma ray burst and a kilonova associated with this event, which allowed the unique identification of the host galaxy, NGC 4993, and hence the recessional velocity (redshift) of the host. The gravitational waves provide a measurement of the luminosity distance of the source. The rate of expansion of the Universe as a function of distance is a key observable for constraining cosmological parameters. The relationship is linear at low distances and the constant of proportionality is called the *Hubble constant*,

$$v = cz = H_0 d,$$

where v is the recessional velocity due to the expansion of the Universe, z is the corresponding redshift, H_0 is the Hubble constant and d is the luminosity distance. At low distance/redshift, the *peculiar velocity* of individual galaxies, relative to the overall expansion of the Universe (the "Hubble flow") is significant and so the observed recessional velocity, v_r, must be corrected by writing $v_r = H_0 d + v_p$. Observations of galaxies provide an estimate of the smoothed peculiar velocity field, $\langle v_p \rangle$. We are interested in inferring the value of the Hubble constant and build a hierarchical model as follows.

- The observed gravitational wave data, x_{GW}, depends on the waveform of the source, which in turn depends on the source parameters. Most of these are not of interest, denoted $\vec{\lambda}$, and so we can marginalise them out, but we treat distance d and inclination, ι, separately

  $$p(x_{\mathrm{GW}} \mid d, \cos\iota) = \int p(x_{\mathrm{GW}} \mid d, \cos\iota, \vec{\lambda})\, p(\vec{\lambda}) \mathrm{d}\vec{\lambda}. \tag{77}$$

- The measured recessional velocity, v_r, depends on the true recessional velocity, which depends on the peculiar velocity, v_p, and the Hubble redshift, $H_0 d$. Representing the electromagnetic measurement uncertainty as a Normal distribution we have

  $$p\left(v_r \mid d, v_p, H_0\right) = N\left[v_p + H_0 d, \sigma_{v_r}^2\right] (v_r) \tag{78}$$

- The measured smoothed peculiar velocity field at the location of the host galaxy depends on the true peculiar velocity there (and perhaps also on other quantities, but we suppress other dependencies here)

$$p\left(\langle v_p\rangle \mid v_p\right) = N\left[v_p, \sigma^2_{v_p}\right]\left(\langle v_p\rangle\right). \tag{79}$$

- The combined likelihood for the observations of $x_{\rm GW}$, $\langle v_p\rangle$ and v_r is

$$p(x_{\rm GW}, v_r, \langle v_p\rangle \mid d, \cos\iota, v_p, H_0) = \\ \frac{1}{\mathcal{N}_s(H_0)} p(x_{\rm GW} \mid d, \cos\iota)\, p(v_r \mid d, v_p, H_0)\, p(\langle v_p\rangle \mid v_p). \tag{80}$$

  Here the factor $\mathcal{N}_s(H_0)$ is the selection effects factor discussed earlier, which corrects for the fact that we only analyse events that exceed some threshold in the gravitational wave detector

$$\mathcal{N}_{\rm s}(H_0) = \int_{\rm detectable} {\rm d}\vec{\lambda}\, {\rm d}d\, {\rm d}v_p\, {\rm d}\cos\iota\, {\rm d}x_{\rm GW}\, {\rm d}v_r\, {\rm d}\langle v_p\rangle \\ \times \Big[p(x_{\rm GW} \mid d, \cos\iota, \vec{\lambda})\, p(v_r \mid d, v_p, H_0) \\ \times p(\langle v_p\rangle \mid v_p)\, p(\vec{\lambda})\, p(d)\, p(v_p)\, p(\cos\iota)\Big]\,, \tag{81}$$

  At the time of GW170817 the horizon for detection of binary neutron stars by the LIGO/Virgo detectors was much smaller ($\sim$ 100Mpc) than the distance to which the kilonova radiation could have been confidently observed ($\sim$ 400Mpc). This means that gravitational wave selection effects were dominant. As these depend directly on the luminosity distance, the dependence on H_0 is a higher order correction and so the selection function was approximately independent of H_0. A correct treatment of election effects will become increasingly important as the LIGO horizon increases in the future.

- We define priors on H_0, d, v_p and $\cos\iota$. These are independent and so we write down a product prior

$$p(d, \cos\iota, v_p, H_0) = p(d)p(\cos\iota)p(v_p)p(H_0).$$

  We use flat priors on $\cos\iota$ and v_p, a volumetric prior on d, $p(d) \propto {\rm d}V_c/{\rm d}d$, where V_c is the comoving volume. We leave $p(H_0)$ unspecified, but note that the analysis in Abbott et al. (2017) used a scale-invariant prior $p(H_0) \propto 1/H_0$.

- We have now fully specified the hierarchical model. A graphical representation of this model is given in Figure 9. The posterior can now be found as

$$p(H_0, d, \cos\iota, v_p \mid x_{\rm GW}, v_r, \langle v_p\rangle) \\ \propto \frac{p(H_0)}{\mathcal{N}_{\rm s}(H_0)} p(x_{\rm GW} \mid d, \cos\iota)\, p(v_r \mid d, v_p, H_0) \\ \times\; p(\langle v_p\rangle \mid v_p)\, p(d)\, p(v_p)\, p(\cos\iota), \tag{82}$$

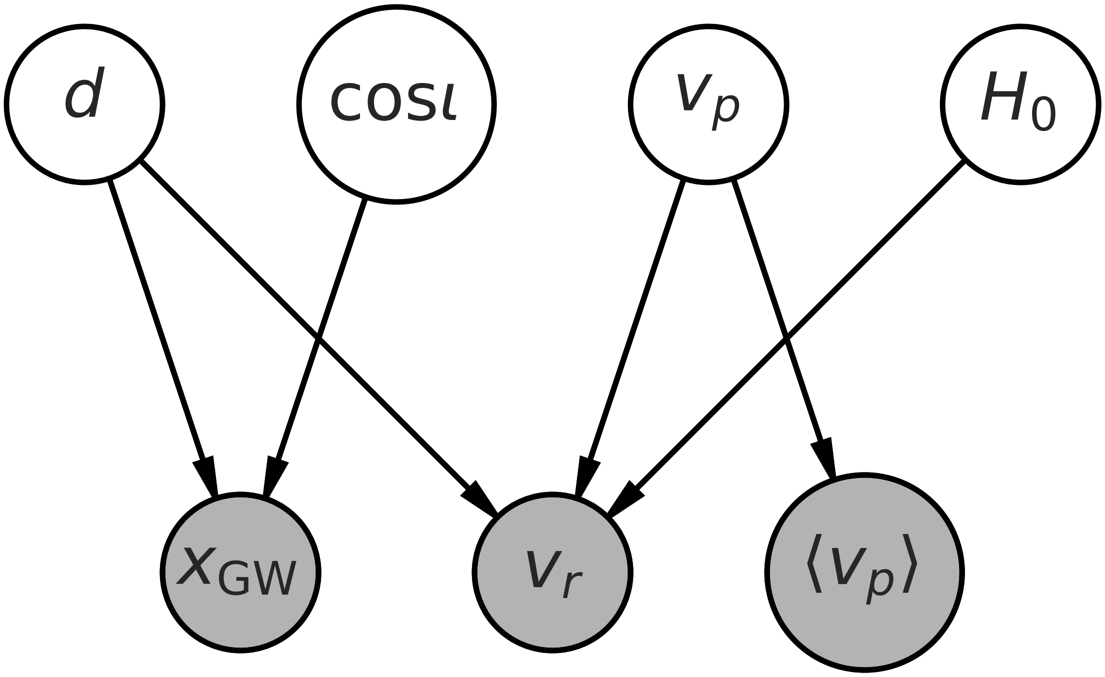


Figure 9: Graphical model for the Hubble constant measurement with gravitational wave observations of binary neutron stars. Figure reproduced from Abbott et al., *Nature Lett.* **551** 85 (2017).

- This posterior can be marginalised over d, $\cos\iota$ and v_p to give

$$p(H_0 \mid x_{\rm GW}, v_r, \langle v_p\rangle) \propto \frac{p(H_0)}{\mathcal{N}_s(H_0)} \int {\rm d}d\, {\rm d}v_p\, {\rm dcos}\,\iota \\ \times p(x_{\rm GW} \mid d, \cos\iota)\, p(v_r \mid d, v_p, H_0) \\ \times p(\langle v_p\rangle \mid v_p)\, p(d)\, p(v_p)\, p(\cos\iota)\,. \quad (83)$$

  This marginalised posterior is shown in Figure 10.

- If we make subsequent observations of binary neutron star mergers with counterparts, indexed by a superscript $i = 1, \ldots, N$, we can combine these

$$p(H_0 \mid \{x^i_{\rm GW}, v^i_r, \langle v_p\rangle^i\}) \propto \frac{p(H_0)}{\mathcal{N}^N_s(H_0)} \prod_{i=1}^{N} \Big[\int {\rm d}d\, {\rm d}v_p\, {\rm dcos}\,\iota \\ \times p(x^i_{\rm GW} \mid d, \cos\iota)\, p(v^i_r \mid d, v_p, H_0) \\ \times p(\langle v_p\rangle^i \mid v_p)\, p(d)\, p(v_p)\, p(\cos\iota)\Big]\,. \quad (84)$$

  Note that, as in the previous example, the prior on the common hyperparameters, $p(H_0)$, occurs only once. The selection effect correction appears once for every observation.

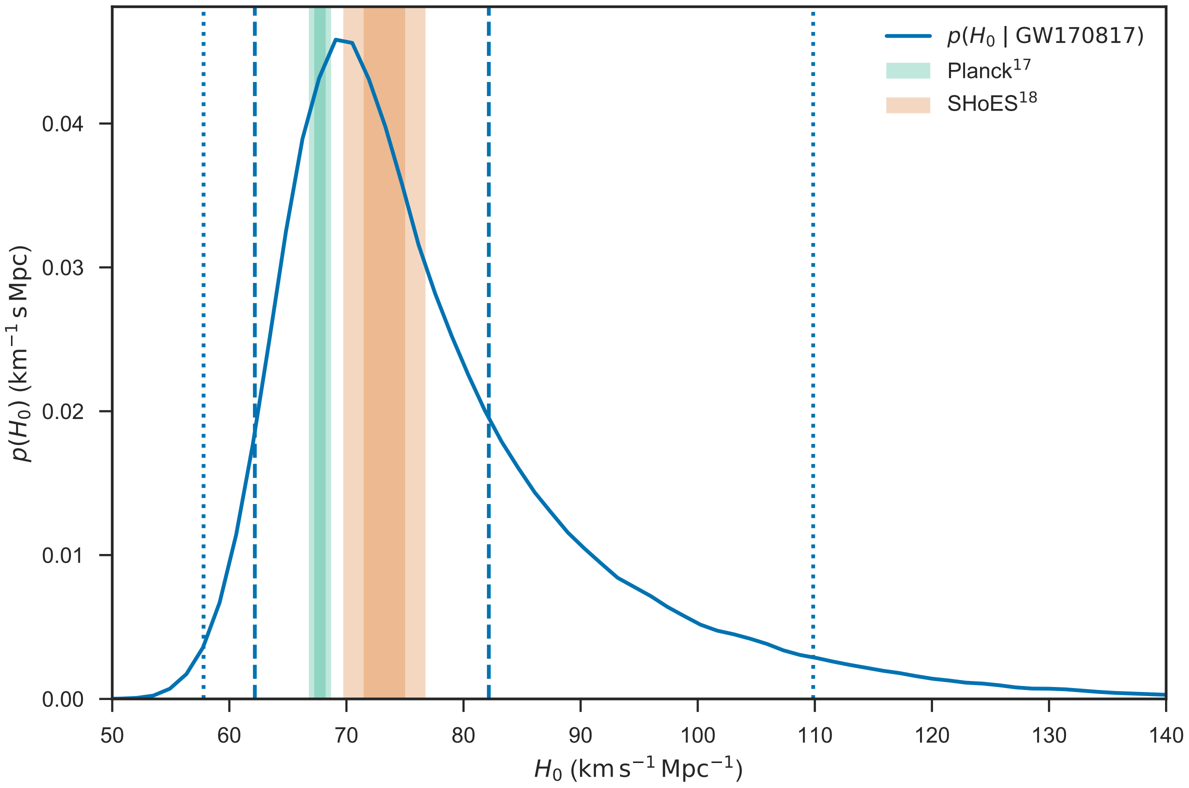


Figure 10: Posterior on the Hubble constant derived from GW170817. Figure reproduced from Abbott et al., *Nature Lett.* **551** 85 (2017).

5 Bayesian Sampling

As emphasised before, the output of Bayesian inference is a posterior probability distribution that encodes our state of knowledge about the parameters of the model based on the observed data and prior information. In certain contexts, for example when using conjugate models, the posterior can be written down in a closed analytic form and used directly for subsequent computation of derived quantities of interest. However, most often the posterior is not known in closed form. There are three approaches to inference in such situations. One is to use a Normal approximation to the posterior, the second is to use brute force integration methods and the third is to draw a set of representative samples from the posterior for us in Monte Carlo integration over the posterior.

5.1 Posterior computation: Bayesian Central Limit Theorem

The **Bayesian Central Limit Theorem** can be used to approximate posteriors, in the limit that the number of observations, $n \to \infty$. Suppose that we have samples $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p(x \mid \boldsymbol{\theta})$ and that the prior, $p(\boldsymbol{\theta})$, and likelihood, $p(x|\boldsymbol{\theta})$, are both twice differentiable near $\widehat{\boldsymbol{\theta}}_{\text{post}}$, the location of the peak of the posterior distribution. Then, for $n \to \infty$, we can approximate

$$p(\boldsymbol{\theta} \mid \mathbf{x}) \sim \mathrm{N}\left(\widehat{\boldsymbol{\theta}}_{\text{post}}, [I^{\text{post}}(\boldsymbol{\theta}, \mathbf{x})]^{-1}\right)$$

where

$$I^{\text{post}}(\boldsymbol{\theta}, \mathbf{x}) = -\left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log p(\boldsymbol{\theta} \mid \mathbf{x})\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{post}}} .$$

The Bayesian central limit theorem follows from the usual central limit theorem. It used to be widely used due to the computational cost of generating posterior samples. However, it relies on the number of observations being large, which is often difficult to ensure in practice. Therefore, its use is no longer so widespread since computers are now sufficiently powerful to enable the generation of large numbers of posterior samples relatively cheaply.

5.2 Posterior computation: numerical integration

In low numbers of dimensions, posterior integrals can be computed using standard numerical integration techniques. There is a large literature on approximating integrals in various ways. The simplest is a grid approach, where the posterior is evaluated at a set of regularly spaced points in the space of waveform parameters. This can be thought of as a type of sampling approximation, where the samples are on a uniform grid. Direct integration rapidly becomes prohibitively expensive as the dimensionality of the model parameter space increases. In addition, it can be inefficient, if the posterior has relatively compact support within the space of allowed values, since many of the grid points will be in regions with low posterior weight.

5.3 Posterior computation: direct sampling methods

As discussed before, sampling methods attempt to generate a set $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$, from the posterior, which can be used to approximate integrals over the posterior

$$\int f(\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathrm{x})\, \mathrm{d}\boldsymbol{\theta} \approx \frac{1}{M}\sum_{i=1}^{M} f(\boldsymbol{\theta}_i).$$

Sampling methods can be **direct** or **stochastic**. Direct methods draw samples directly (or nearly directly) from the target probability distribution. Stochastic methods use **Markov chain Monte Carlo** methods to generate a sequence of samples that are drawn form the target distribution.

5.3.1 Method of inversion

The method of inversion is a simple application of the probability integral transformation. If we denote by F the cumulative distribution function of some random variable X, then the random variable $F(X)$ follows a $U[0,1]$ distribution. Therefore, if we can analytically compute the inverse of the cumulative distribution function, we can generate samples form X by generating samples from a uniform distribution. If

$$F(x) = \mathbb{P}(X \leq x)$$

and it has inverse F^{-1} then the algorithm is simply

1. Generate $u \sim U[0,1]$.
2. Compute $x = F^{-1}(u)$.

Example: exponential distribution with parameter r Suppose we want to draw $X \sim \mathrm{Exp}(r)$. The pdf of the exponential distribution is

$$p(x|r) = r\exp(-rx)$$

which has cumulative density function

$$F(X) = \int_0^X r\exp(-rx)\mathrm{d}x = 1 - \exp(-rX).$$

The inverse can be found as

$$u = F(x) \quad \Rightarrow \quad x = F^{-1}(u) = -\frac{1}{r}\ln(1-u).$$

Samples generated by applying this inverse to $U[0,1]$ samples are shown in Figure 11.

5.3.2 Rejection sampling

Rejection sampling draws samples from a distribution that can be directly sampled and then discards a subset of them that do not match the desired distribution. The simplest rejection sampling algorithm draws uniform samples from a box that encloses the distribution. Suppose that we want to draw samples $\theta_1, \ldots, \theta_n$ from a probability distribution with pdf $p(\theta)$ and that the pdf has compact support, so $p(\theta) = 0$ if $\theta \notin [a,b]$. Suppose additionally that the pdf at the mode of the probability distribution is $M = \max[p(\theta)]$. Rejection sampling proceeds as follows

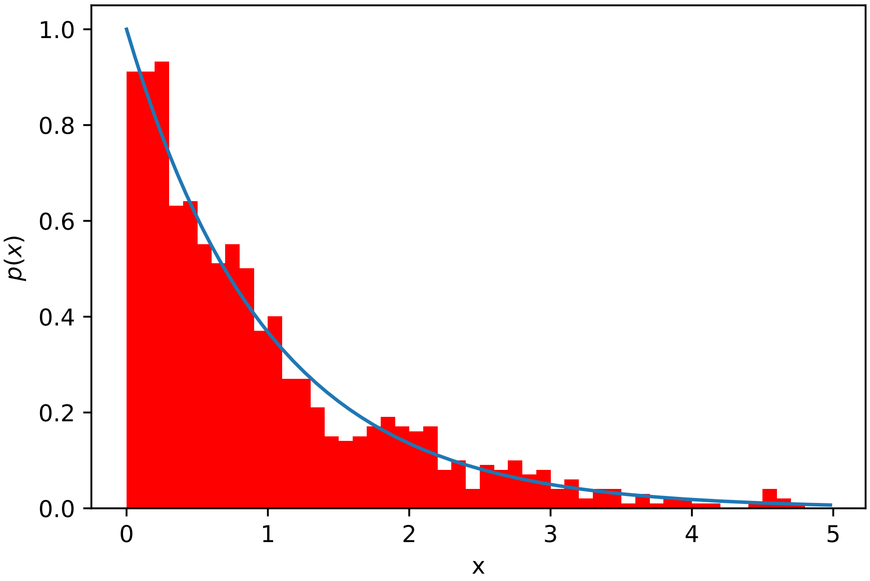


Figure 11: Histogram of samples drawn from the Exp(1) distribution using the method of inversion. The pdf of the exponential distribution is shown as a line for comparison.

1. Draw $\theta \sim U[a, b]$.
2. Draw $y \sim U[0, M]$.
3. If $y \leq p(\theta)$, accept θ as a sample from $p(\theta)$. Otherwise return to step 1.

Example: beta distribution We want to draw samples from a Beta$(3, 2)$ distribution. This has compact support on the interval $[0, 1]$ and the maximum value of the pdf is $M = 16/9$ (EXERCISE). In Figure 12 we illustrate this procedure by indicating which of the first 50 samples drawn in this way are rejected or accepted. In Figure 15 we show a histogram of the accepted samples in the first 1000 draws, which illustrates that the distribution of samples does follow the Beta$(3, 2)$ distribution as desired.

The box rejection sampling procedure does not work at all when the support of the target distribution is unbounded. In addition, it can be very inefficient for compact distributions with long tails. An alternative approach is to draw samples from an easy-to-sample distribution, $g(\theta)$, that is similar to the target distribution $p(\theta)$. First we find a number M such that $Mg(\theta) \geq p(\theta)\ \forall \theta$, i.e., we require $Mg(\theta)$ to contain the target distribution. The algorithm is then

1. Draw $\theta \sim g(\theta)$.
2. Draw $y \sim U[0, 1]$.
3. If $y \leq p(\theta)/(Mg(\theta))$, accept θ as a sample from $p(\theta)$. Otherwise return to step 1.

Trial samples are taken uniformly from within the region between the curve $Mg(\theta)$ and the θ axis. Samples that fall in the region between $p(\theta)$ and $Mg(\theta)$ are rejected. Therefore we

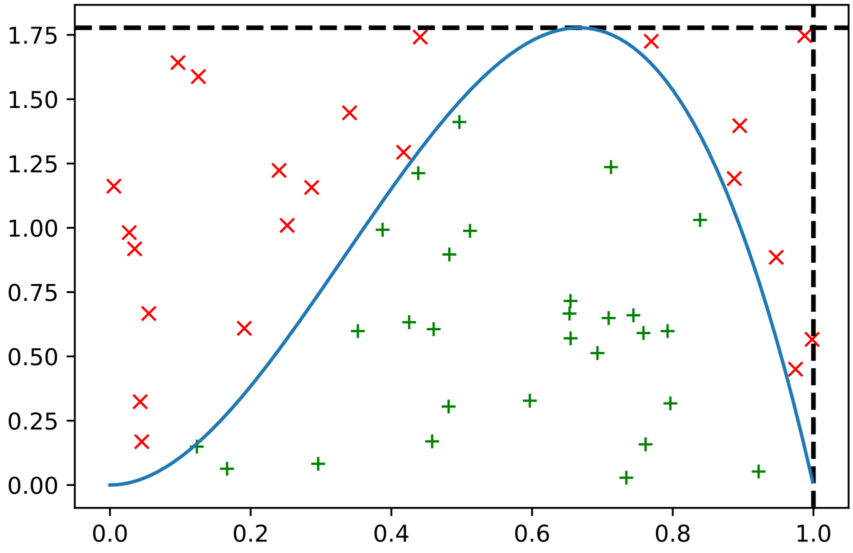


Figure 12: Accepted (green plusses) and rejected (red crosses) samples in the first 50 draws of the rejection sampling algorithm used to simulate the Beta$(3, 2)$ distribution. Only samples that lie within the target pdf are accepted.

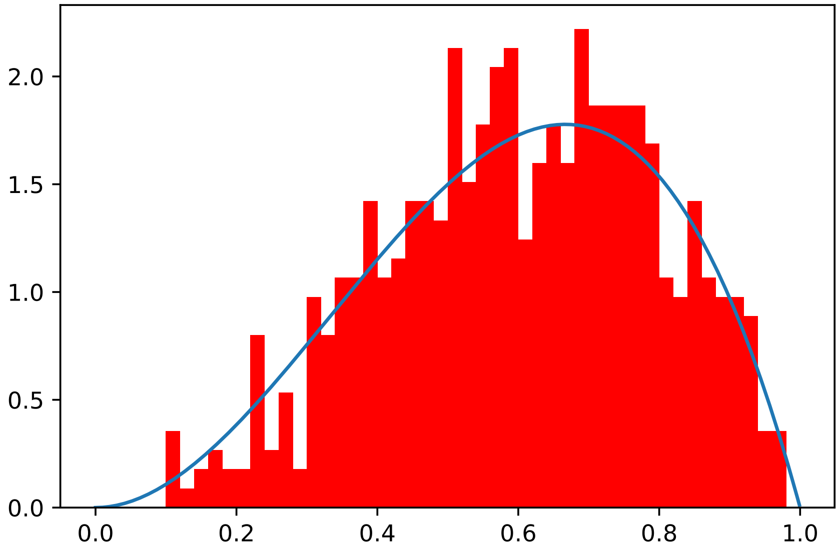


Figure 13: Histogram of the accepted samples in 1000 iterations of the rejection sampling algorithm. We compare the distribution to Beta$(3, 2)$, which is the target distribution.

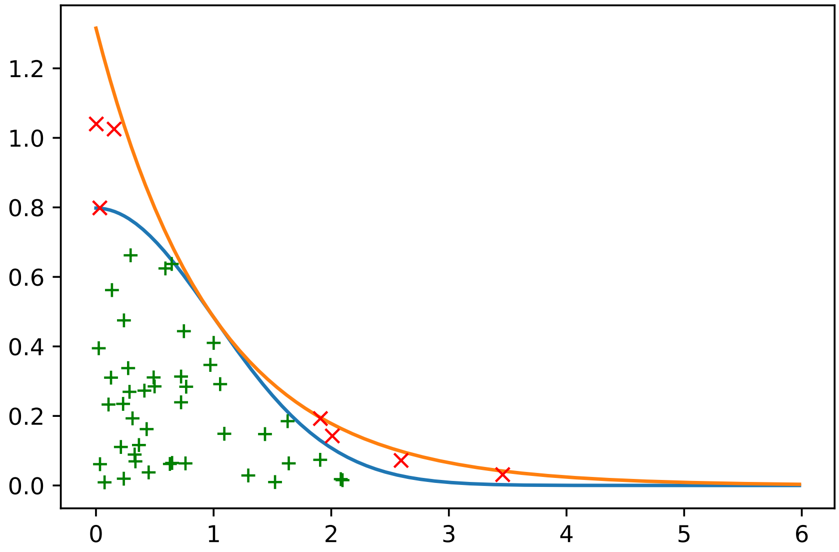


Figure 14: Accepted (green plusses) and rejected (red crosses) samples in the first 50 draws of the rejection sampling algorithm to simulate draws from a half-Normal distribution. Only samples that lie within the target pdf are accepted.

make the efficiency (i.e., the fraction of samples that are accepted) as large as possible by making the choice

$$M = \sup_\theta \left(\frac{p(\theta)}{g(\theta)} \right) .$$

Example: half-Normal distribution We want to draw samples from the half-Normal distribution with pdf

$$p(\theta) = \begin{cases} \sqrt{\frac{2}{\pi}} \mathrm{e}^{-\frac{\theta^2}{2}} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases} .$$

We will take $g(\theta) = \exp(-\theta)$, i.e., the exponential distribution with rate 1. We find M from

$$M = \sup_\theta \left(\frac{p(\theta)}{g(\theta)} \right) = \sup_{\theta > 0} \left(\sqrt{\frac{2}{\pi}} \exp\left[-\frac{1}{2}(\theta - 1)^2 + \frac{1}{2} \right] \right) = \sqrt{\frac{2}{\pi}} \mathrm{e}^{\frac{1}{2}} .$$

In Figure 14 we show the samples accepted and rejected during the first 50 iterations of the algorithm, and in Figure **??** we show a histogram of the accepted samples during 1000 iterations of the algorithm. We see that the histogram is correctly approximating the desired distribution.

5.3.3 Importance sampling

Rejection sampling can be effective and easy to implement, but it is not always possible to find an easy-to-sample target distribution that closely matches the target distribution. Additionally effort is wasted drawing samples and evaluating the posterior at points which

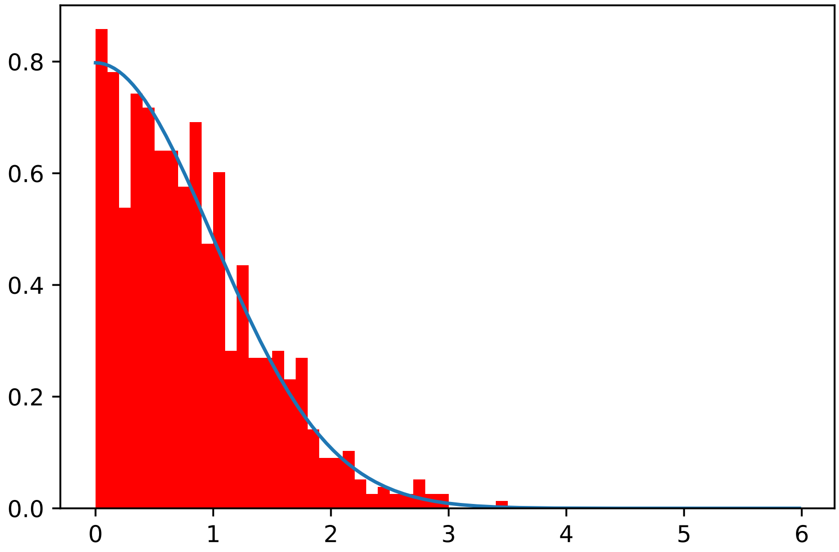


Figure 15: Histogram of the accepted samples in 1000 iterations of the rejection sampling algorithm. We compare the distribution to the target distribution, which in this case is a half-Normal distribution with mean 0 and variance 1.

are subsequently discarded as rejected samples. **Importance sampling** attempts to address the latter problem by using all samples.

Importance sampling uses an easy-to-sample reference distribution $g(\theta)$ as before, but now this is not rescaled, the only stipulation is that the support is common to that of the target distribution, i.e., if $p(\theta) > 0$ then $g(\theta) > 0$. No samples are discarded. Instead the samples are defined **importance weights** via

$$w_i = \frac{p(\theta)}{g(\theta)}$$

and integrals over the target distribution are approximated by weighted averages over the samples

$$\int f(\theta)p(\theta)\mathrm{d}\theta \approx \frac{1}{M}\sum_{i=1}^{M} w_i f(\theta_i).$$

It is straightforward to see that

$$\mathbb{E}_g(w_i f(\theta_i)) = \int w(\theta)f(\theta)g(\theta)\mathrm{d}\theta = \int \frac{p(\theta)}{g(\theta)}f(\theta)g(\theta)\mathrm{d}\theta = \int f(\theta)p(\theta)\mathrm{d}\theta = \mathbb{E}_p(f(\theta))$$

so the importance sampling estimate is unbiased. However

$$\begin{aligned}\mathrm{var}_g(w_i f(\theta_i)) &= \int w^2(\theta)f^2(\theta)g(\theta)\mathrm{d}\theta - [\mathbb{E}_p(f(\theta))]^2 = \int \frac{p(\theta)}{g(\theta)}f^2(\theta)p(\theta)\mathrm{d}\theta - [\mathbb{E}_p(f(\theta))]^2 \\ &= \mathbb{E}_p\left(\frac{p(\theta)}{g(\theta)}f^2(\theta)\right) - [\mathbb{E}_p(f(\theta))]^2. \end{aligned} \quad (85)$$

We see that the importance sampling estimate can suffer from high variance if $g(\theta)$ is much smaller than $p(\theta)$ in regions where the function of interest has significant support.

Note that the above assumes that the normalisation of the target distribution is known, but this is not always the case when sampling from posterior distributions due to the difficulty of computing the Bayesian evidence. If the posterior is not normalised the weights can be renormalised as

$$\tilde{w_i} = \frac{w_i}{\sum_{j=1}^{M} w_j}.$$

The results on the mean and variance are now only approximate, but are valid asymptotically.

Example: Cauchy distribution Suppose we have a standard Cauchy distribution with pdf

$$p(\theta) = \frac{1}{\pi(1+\theta^2)}$$

and want to compute $\mathbb{P}(\theta > 2)$. We can sample from the distribution $g(\theta) = 2/\theta^2 \mathbb{I}(\theta > 2)$ using the method of inversion. This has the same support as the portion of $p(\theta)$ of interest. We define the importance weights

$$w_i = \frac{\theta_i^2}{2\pi(1+\theta_i^2)}$$

and then compute

$$\hat{p}_{>2} = \frac{1}{M}\sum_{i=1}^{M} w_i$$

since we are interested in $\mathbb{P}(\theta > 2)$ which is the integral of $\mathbb{I}(\theta > 2)$, but this equal to 1 throughout the region where $g(\theta)$ has support. Note that in this case it would be wrong to renormalise the weights since then we would compute the probability as 1. As an exercise, verify that using the above weights in the usual sampling estimate gives the expected result.

In Figure 16 we show the convergence of the importance sampling estimate of $\mathbb{P}(\theta > 2)$ as a function of the number of importance samples. We see that it converges much faster than if we used Monte Carlo draws from the Cauchy distribution itself. The correct probability is $\pi/2 - \tan^{-1}(2)/\pi = 0.14758$.

5.3.4 Sampling importance resampling

Sampling importance resampling is a simple extension of importance sampling that uses the importance samples to generate samples approximately from the target distribution. Given M importance samples, $\{\theta_1, \ldots, \theta_M\}$, the importance weights are computed and normalised as described above. Then M samples, $\{\phi_1, \ldots, \phi_M\}$ are drawn, with replacement, from the original set using the normalised weights as probabilities. Integrals over the target distribution can then be approximated by

$$\int f(\theta)p(\theta)\mathrm{d}\theta \approx \frac{1}{M}\sum_{i=1}^{M} f(\phi_i).$$

Sampling importance resampling is a form of **particle filtering**. One problem that it can suffer form is **particle depletion**, where a small number of samples carry the majority of

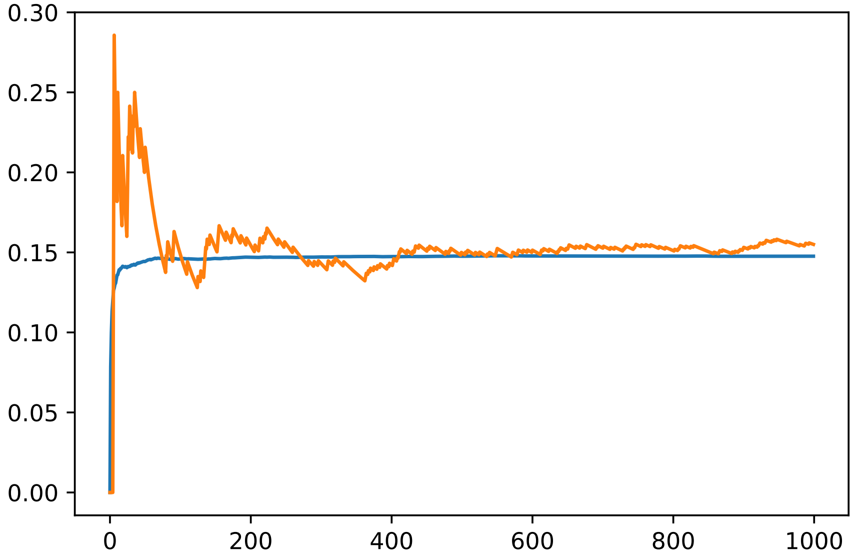


Figure 16: Importance sampling estimate of $\mathbb{P}(\theta > 2)$ for the standard Cauchy distribution as a function of number of samples (blue line), compared to Monte Carlo estimate using direct samples from the Cauchy distribution (yellow line).

the weight and therefore only a small number of points are represented repeatedly in the final data set. Particle depletion leads to poor estimates of derived quantities.

Example: Cauchy distribution We use sampling importance resampling to generate samples from the Cauchy distribution with $\theta > 2$ using the samples generated for the example in the previous section. A histogram of these values is shown in Figure 17, where they are compared to the target distribution, which is a truncated Cauchy distribution.

5.4 Posterior computation: Markov chain Monte Carlo

Direct sampling methods suffer form the problem of dimensionality. They are typically easy to implement in one dimension, but become increasingly challenging, inefficient or impossible to implement as the number of dimensions increases. In higher dimensions it is more common to use stochastic methods, in which a sequence of samples is constructed that has a distribution that follows the target distribution. Typically this is done using Markov chain Monte Carlo algorithms.

A **Markov Chain** is a sequence of random numbers, $\theta^1, \theta^2 \ldots$, such that the value of θ^{n+1} depends only on the previous values, θ^n, and not on earlier numbers in the sequence. A Markov chain can be simulated using a **transition kernel**, $\mathcal{K}(\theta^{n+1}|\theta^n)$, which is a conditional probability distribution for θ^{n+1} given the value of θ^n. The transition kernel uniquely defines the Markov chain. If we assume the Markov chain is **aperiodic** and **irreducible** then the distribution of samples in the Markov chain will converge to a **stationary distribution**, which is independent of the initial starting state of the chain. In Bayesian inference, the goal is to construct a Markov chain such that the stationary distribution is the posterior distribution, $p(\boldsymbol{\theta}|\mathbf{x})$.

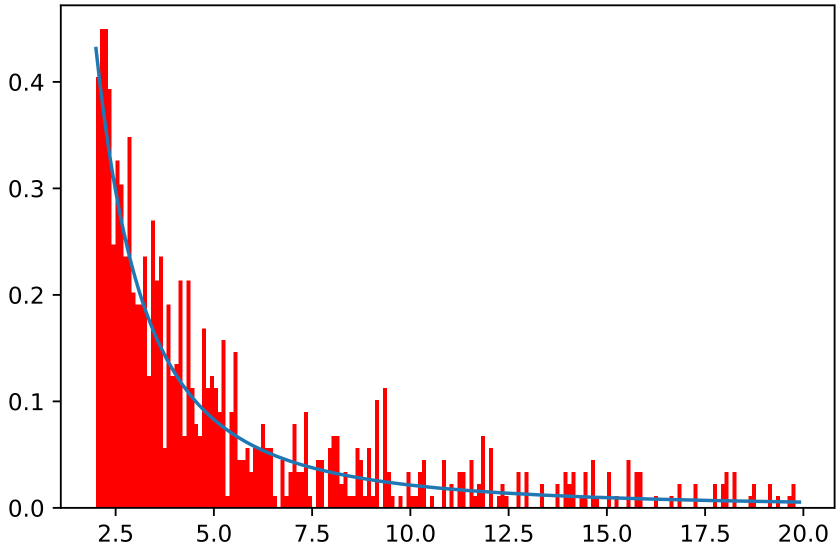


Figure 17: Histogram of 1000 sampling importance resampling samples for the Cauchy distribution in the region $\theta > 2$. These were generated from the importance samples constructed in the example in the last section. The line shows the expected distribution, which is the truncated standard Cauchy distribution.

A Markov chain with transition kernel $\mathcal{K}(\theta^{n+1}|\theta^n)$ is said to satisfy **detailed balance** for a distribution $\pi(\theta)$ if

$$\pi(\theta)\mathcal{K}(\phi|\theta) = \pi(\phi)\mathcal{K}(\theta|\phi) \;\forall\, \phi, \theta,$$

in which case $\pi(\theta)$ is the stationary distribution of the Markov chain. Enforcing detailed balance in the Markov chain, for $\pi(\theta) = p(\theta|\mathbf{x})$, will ensure we generate samples from the posterior distribution.

There are two widely used approaches to construct Markov chains satisfying detailed balance with a particular stationary distribution — Gibbs sampling and the Metropolis-Hastings algorithm.

5.4.1 Gibbs Sampling

Gibbs sampling for multi-variate probability distributions works by sampling sequentially from full conditional distributions on each parameter given the current state of the other parameters. Algorithmically it works as follows. We suppose that the distribution of interest, $p(\boldsymbol{\theta}|\mathbf{x})$, depends on a multi-dimensional parameter vector, $\boldsymbol{(}\theta\boldsymbol{)} = (\theta_1, \theta_2, \ldots, \theta_p)$. We use $\boldsymbol{(}\theta\boldsymbol{)}^k$, θ_i^k to denote the value of the full parameter vector and its i'th component at iteration k of the algorithm. We denote by $\boldsymbol{\theta}_{(i)}$ the vector of all parameter values except the i'th and use $p(\theta_i|\boldsymbol{\theta}_{(i)}, \mathbf{x})$ to denote the full conditional distribution of θ_i, given the values of all the other components and the data. If the value of the Markov chain at step t is $\boldsymbol{\theta}^t$, then the value at step $t+1$ is obtained via

- Sample θ_1^{t+1} from $p(\theta_1|\theta_2^t, \theta_3^t, \ldots, \theta_p^t, \mathbf{x})$.

- Sample θ_2^{t+1} from $p(\theta_2|\theta_1^{t+1}, \theta_3^t, \ldots, \theta_p^t, \mathbf{x})$.

- . . . .
- Sample θ_i^{t+1} from $p(\theta_i|\theta_j^{t+1}$ for $j < i$ and θ_j^t for $j > i, \mathbf{x})$.
- . . . .
- Sample θ_p^{t+1} from $p(\theta_p|\theta_1^{t+1}, \dots, \theta_{p-1}^{t+1}, \mathbf{x})$.

This set of sequential updates is repeated at each iteration of the algorithm to generate a set of samples from the target distribution.

The transition kernel in Gibbs sampling is

$$\mathcal{K}_G(\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^t) = \prod_{i=1}^{k} p(\theta_i|\theta_j^{t+1} \text{ for } j < i \text{ and } \theta_j^t \text{ for } j > i, \mathbf{x})$$

which satisfies detailed balance with target distribution $p(\boldsymbol{\theta}|\mathbf{x})$.

5.4.2 Metropolis-Hastings algorithm

In the Metropolis-Hastings algorithm all the parameters of the model are typically updated simultaneously. This is achieved using a **proposal distribution**, $q(\boldsymbol{\phi}|\boldsymbol{\theta})$, to propose a new point $\boldsymbol{\phi}$, given the current parameter values $\boldsymbol{\theta}$. The algorithm is as follows

1. Initialise $\boldsymbol{\theta}^0$ by drawing from a distribution of starting vlaues (often the prior can be used for this).
2. At step t:
   (a) Propose a new point $\boldsymbol{\phi} \sim q(\boldsymbol{\phi}|\boldsymbol{\theta}^{t-1})$.
   (b) Compute the **acceptance probability**
   $$\alpha = \min\left(1, \frac{p(\boldsymbol{\phi}|\mathbf{x})q(\boldsymbol{\theta}^{t-1}|\boldsymbol{\phi})}{p(\boldsymbol{\theta}^{\boldsymbol{t-1}}|\mathbf{x})q(\boldsymbol{\phi}|\boldsymbol{\theta}^{t-1})}\right).$$
   (c) Draw $u \sim U[0,1]$. If $u < \alpha$, set $\boldsymbol{\theta}^t = \boldsymbol{\phi}$, otherwise set $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1}$.
3. Repeat until the desired number of iterations, T, have been completed.

the initial version of this algorithm, due to Metropolis, used symmetric proposal distributions and so that factor cancels out of the acceptance probability. A subsequent paper by Metropolis and Hastings generalised the result to non-symmetric proposals.

It can be readily verified in this case as well that the Markov chain constructed in this way satisfies detailed balance with target distribution equal to the posterior $p(\boldsymbol{\theta}|\mathbf{x})$.

There are a few special cases of the Metropolis-Hastings algorithm

- **The Metropolis Algorithm** This is the case described above where the proposal is symmetric, $q(\boldsymbol{\phi}|\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\boldsymbol{\phi})$, and the acceptance probability reduces to
$$\alpha = \min\left(1, \frac{p(\boldsymbol{\phi}|\mathbf{x})}{p(\boldsymbol{\theta}^{\boldsymbol{t-1}}|\mathbf{x})}\right).$$

- **Random Walk Metropolis** If we use $q(\boldsymbol{\phi}|\boldsymbol{\theta}) = f(\boldsymbol{\theta} - \boldsymbol{\phi})$, with f some function satisfying $f(\mathbf{y}) = f(-\mathbf{y})$, then the kernel driving the chain is a random walk. This is a symmetric proposal and so the accpetance probability is as in the Metropolis Algorithm above.

- **The Independence Sampler** If we take $q(\boldsymbol{\phi}|\boldsymbol{\theta}) = f(\boldsymbol{\phi})$, the candidate value is independent of the current value. The acceptance probability is

$$\alpha = \min\left(1, \frac{w(\boldsymbol{\phi})}{w(\boldsymbol{\theta})}\right)$$

  where $w(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{x})/f(\boldsymbol{\theta})$.

- **Single-updates** Individual parameters of the parameter vector can be updated sequentially in the Metropolis-Hastings algorithm in the same way they are during the Gibbs sampling algorithm. At step t we sequentially propose updates, ϕ_j, to each component, θ_j, of the parameter vector in turn. After updating parameter j, the new parameter vector is $(\theta_1^{t+1}, \ldots \theta_{j-1}^{t+1}, \theta_j,^{t+1} \theta_{j+1}^{t}, \ldots, \theta_p^t)$. The new value, θ_j^{t+1} is chosen by the algorithm

  1. Propose a new candidate value $\phi_j \sim q(\phi_j|\boldsymbol{\theta}_j^t)$ and set $\boldsymbol{\phi}_j = (\theta_1^{t+1}, \ldots, \theta_{j-1}^{t+1}, \phi_j, \theta_{j+1}^t, \ldots, \theta_p^t)$.
  2. Evaluate the acceptance probability

$$\alpha = \min(1, A), \quad \text{where} \quad A = \frac{p(\boldsymbol{\phi}_j|\mathbf{x})q(\theta_j^t|\boldsymbol{\phi}_j)}{p(\boldsymbol{\theta}_j^t|\mathbf{x})q(\phi_j|\boldsymbol{\theta}_j^t)} = \frac{p(\phi_j|\boldsymbol{\theta}_{(j)}^t, \mathbf{x})q(\theta_j^t|\boldsymbol{\phi}_j)}{p(\theta_j^t|\boldsymbol{\theta}_{(j)}^t, \mathbf{x})q(\phi_j|\boldsymbol{\theta}_j^t)}$$

  3. Draw $u \sim U[0,1]$. If $u < \alpha$, set $\theta_j^{t+1} = \phi_j$, otherwise set $\theta_j^{t+1} = \theta_j^t$.

5.4.3 MCMC diagnostics

The Markov chain is only guaranteed to converge to the stationary distribution asymptotically so it is natural to ask how many samples are needed before the sample is representative of the posterior. The first issue to address is **burn-in**. A Markov chain retains some memory of its initial state for a number of iterations. If the initial sample is in a region of low probability in the stationary distribution, then the first samples will typically not be very characteristic of the stationary distribution. These initial samples should be discarded and samples only retained after the initial burin-in period used for inference. Typically between a few hundred and a few thousand burn-in samples are required and it can be diagnosed using a **trace plot**, which is a plot of the parameter value in the chain as a function of iteration number. Initially the trace plot will show a trend as the chain moves toward parameter values with high posterior support. Once the chain is sampling properly, the values will oscillate back and forth. This is illustrated in Figure 18. The trace plot allows the burn-in period to be identified and removed, and is also a useful diagnostic of the performance of the algorithm. Chains that are moving back and forth rapidly are sampling well from the posterior.

MCMC samples are used to produce Monte Carlo estimates of parameters of interest. If the samples were independent draws from the posterior then these estimates are unbiased and would have a variance that scales like σ^2/M, where σ^2 is the variance of a single sample and M is the number of samples. This could in principle be used to estimate how many

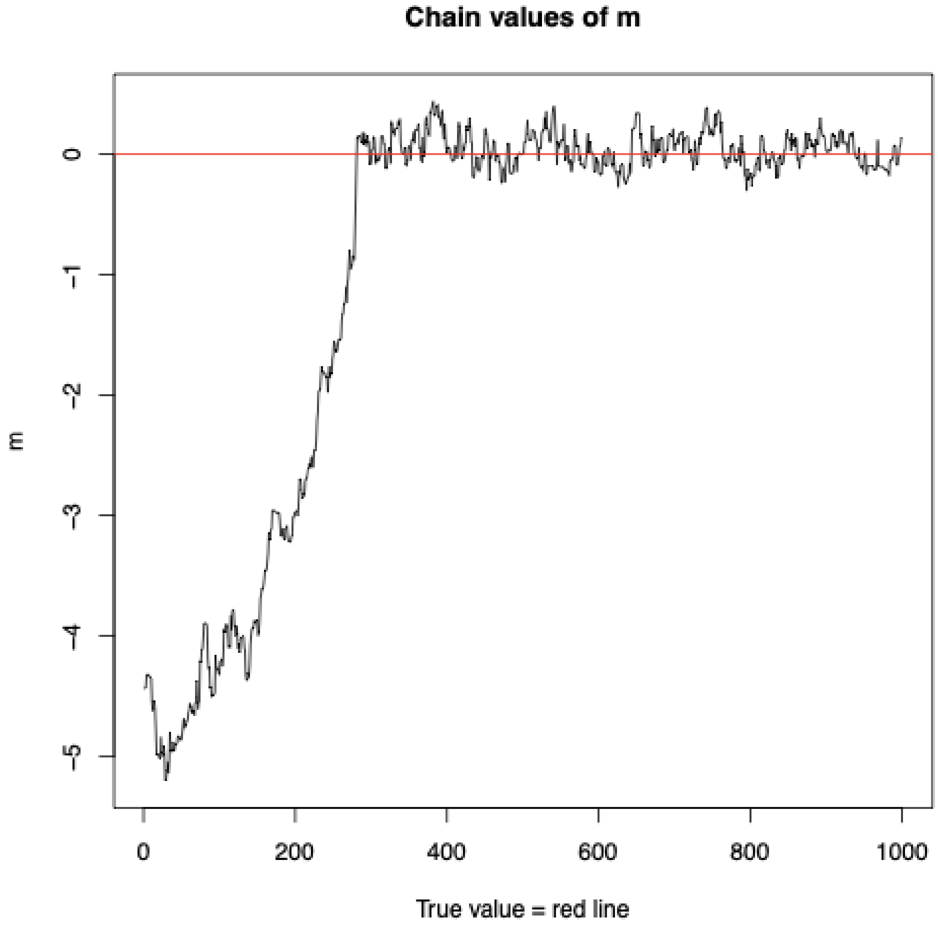


Figure 18: Trace plot for burn-in period of a chain. Initially the chain moves form the starting point to the region of high probability density, so there is a tendency to move in a particular direction. Once the chain reaches the correct region it oscillates back and forth in the region of high posterior support.

samples are needed to achieve a certain target precision on a quantity of interest. However, MCMC samples are not independent. This modifies the variance estimate to

$$\sigma^2 = \mathrm{var}_p(\theta) + 2\sum_{k=2}^{\infty} \mathrm{cov}(\theta^1, \theta^k).$$

This is difficult to compute in practice, so what is usually done is to generate m different chains of length M, estimate the value of the quantity of interest in each one, $\bar{f}_1, \ldots, \bar{f}_m$, compute the value using the pooled samples from all chains, $\bar{f}$, and then construct the *batch means estimate*

$$\hat{\sigma}^2 = \frac{T}{m-1}\sum_{i=1}^{m}(\bar{f}_i - \bar{f})^2.$$

The estimated Monte Carlo error in f is then $\hat{\sigma}^2/n$.

Correlation in MCMC samples can also be estimated using the **autocorrelation function** (ACF). The *lag-k autocorrelation coefficient* or *autocorrelation at lag-k* is $\mathrm{cov}(\theta^i, \theta^{i+k})$ and computed via

$$\rho_k = \frac{\sum_{i=1}^{N-k}(\theta^i - \bar{\theta})(\theta^{i+k} - \bar{\theta})}{\sum_{i=1}^{M}(\theta^i - \bar{\theta})^2}$$

where θ now denotes one parameter of the target distribution, and $\bar{\theta}$ is the mean of that parameter in the chain. Looking at ACF plots is another useful diagnostic of MCMC performance. Examples of good, bad and normal ACF plots are given in Figure 19.

If MCMC chains have very high lags, most likely they are not taking big enough jumps in parameter space and so the size of proposed jumps should be increased. It is typical to monitor **acceptance rates** when using the Metropolis-Hastings algorithm and a target acceptance rate is used to adjust proposed jump sizes. If proposed jumps are too small, the acceptance rate will be high but there will also be high autocorrelation between samples. If the proposed jumps are too large, the acceptance rate will be low, but those samples that are accepted will show very low autocorrelation.Ultimately we care about maximising the rate at which we obtain new independent samples. This can be estimated by tracking the **effective sample size**

$$\mathrm{ESS} = \frac{M}{1 + 2\sum_{k=1}^{\infty}\rho_k}$$

where M is the number of samples in the chain. It has been shown that, under certain assumptions, the optimal rate of obtaining new effective samples is achieved by aiming to have an acceptance rate around 23.4%.

The final diagnostic we will mention here is the use of multiple chains. For complex probability distributions that have many modes it is possible for Markov chains to get stuck sampling from only one of them. Chains starting from different points in parameter space may end up exploring different modes. As a diagnostic of this kind of behaviour, it is good practice to run a handful of runs, starting at different points in parameter space. We can be confident in the final results once the different chains are producing samples that are consistent with one another. This consistency can be quantified using the **Gelman-Rubin statistic**.

Suppose we have m independent chains and have discarded the initial burn-in samples

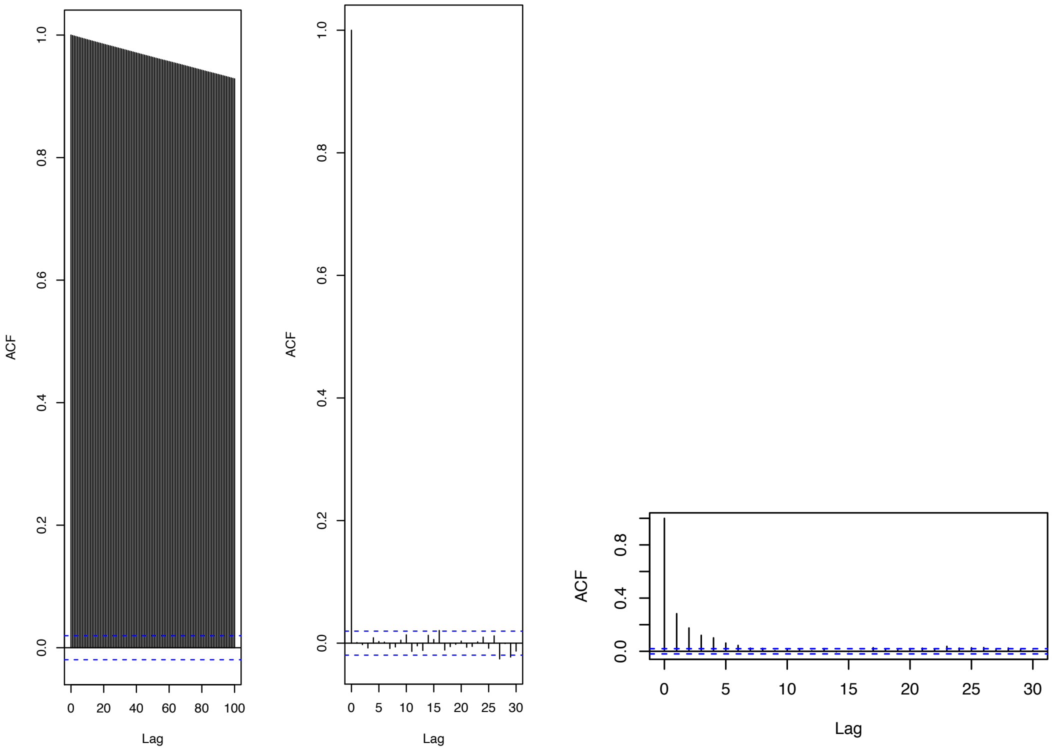


Figure 19: Examples of plots of the autocorrelation function. This should decline to numbers close to 0 for short lags. In the left hand plot, the ACF is still above 0.8 at a lag of 100, indicating highly correlated samples, which is not desirable. In the middle plot we show an ideal example where the ACF is already close to zero at lag of 1, indicating a high level of independence in the samples. The right hand plot is a typical example of MCMC chains that are sampling well. The ACF falls to low values for lags of a few.

to leave chains of length N. We calculate the *within chain variance*

$$W = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{N-1}\sum_{i=1}^{N}(\theta_{ij}-\bar{\theta}_j)^2$$

where θ_{ij} is the i'th sample in the j'th chain. We similarly define the *between chain variance*

$$B = \frac{N}{m-1}\sum_{j=1}^{m}(\bar{\theta}_j-\bar{\bar{\theta}})^2, \quad \text{where } \bar{\bar{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}_j.$$

Note that we are assuming that θ is a one-dimensional parameter, which could be one component of a multi-dimensional parameter vector. The variance in this parameter can be computed as

$$\mathrm{var}(\theta) = \left(1-\frac{1}{N}\right)W + \frac{1}{N}B$$

from which the **potential scale-reduction factor** can be computed

$$\hat{R} = \sqrt{\frac{\mathrm{var}(\theta)}{W}}.$$

Values of R greater than about 1.1 or 1.2 indicate that the chains are not yet converged.

5.4.4 Speeding up MCMC

MCMC can be made faster by a good choice of the proposal distribution. Proposal distributions that are well approximated to the form of the target distribution are to be preferred. As well as tuning the proposal distribution, accelerated convergence can be achieved using **annealing**. The idea of annealing is to transform the posterior surface as

$$p(\boldsymbol{\theta}|\mathbf{x}) \to [p(\boldsymbol{\theta}|\mathbf{x})]^{\beta}, \quad \text{where } \beta = \frac{1}{kT}.$$

As $T \to \infty$ the new distribution becomes flatter and flatter, so the contrast in probabilities between different points is reduced. This means that moves proposed in a Metropolis-Hastings algorithm are more likely to be accepted. Figure 20 shows the effect of the annealing transformation on the probability distribution being sampled as the temperature increases.

There are two common applications of annealing. In **simulated annealing** the temperature is gradually changed as the initial phase of the run progresses, according to some scheme, for example, a linear decrease with iteration number. The idea is that in the early phase the chain explores the parameter space widely and rapidly, identifying areas of higher posterior density. As the temperature decreases the chain gets trapped in a region of high posterior probability, hopefully the primary mode of the distribution. The simulated annealing phase does not produce useful samples, since detailed balance is satisfied, but after the simulated annealing phase, the chain will evolve as normal and return valid samples from the posterior.

The other use of annealing is **parallel tempering**. In parallel tempering, a number of chains are evolved simultaneously at different temperatures. At each iteration, a given chain will update its parameters as normal, but with a certain probability an interchange is proposed, in which the states of two chains (usually neighbouring in temperature) will

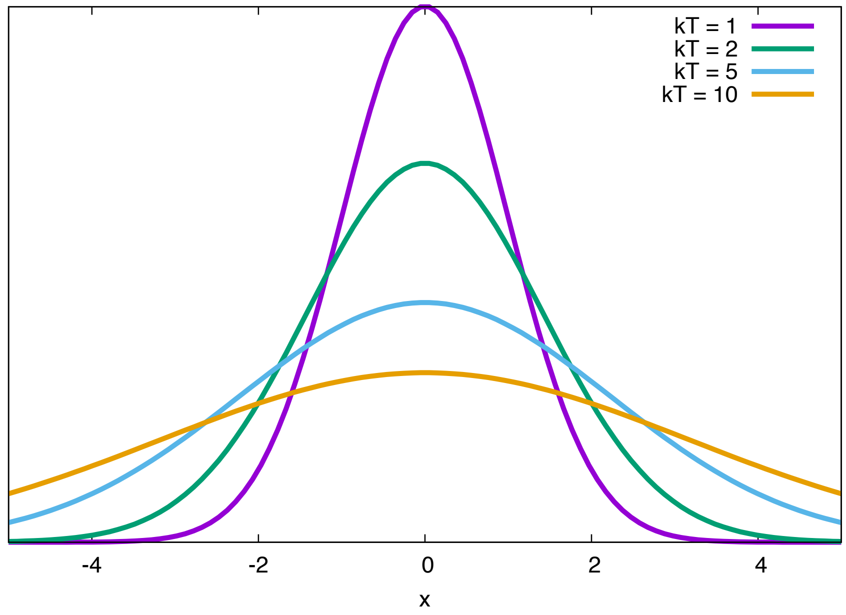


Figure 20: Effect of annealing on the target probability distribution.

be exchanged. If the two chains are labelled i and j, have temperatures T_i and T_j, and current parameter values $\boldsymbol{\theta}^i$ and $\boldsymbol{\theta}^j$, then the appropriate acceptance probability for the swap $\boldsymbol{\theta}^i \leftrightarrow \boldsymbol{\theta}^j$ is

$$\alpha = \min\left(1, \left[\frac{p(\boldsymbol{\theta}^j|\mathbf{x})}{p(\boldsymbol{\theta}^i|\mathbf{x})}\right]^{\frac{1}{T_i}} \left[\frac{p(\boldsymbol{\theta}^i|\mathbf{x})}{p(\boldsymbol{\theta}^j|\mathbf{x})}\right]^{\frac{1}{T_j}}\right).$$

The idea of parallel tempering is that higher posterior density regions of the parameter space that the widely-exploring high temperature chains identify, propagate down to lower temperature chains, which explore them thoroughly. Efficiency is dependent on the difference in the temperatures of neighbouring chains, so the number of chains and their spacing must be tuned for each given problem.

5.5 Posterior computation: variable model dimension

In some circumstances we might be interested in fitting multiple different models to the data simultaneously. the most common situation is when the total number of parameters needed to describe the data is unknown. In a gravitational wave context this arises when the total number of sources present in the data set is unknown, e.g., for the LISA gravitational wave detector. In these circumstances one can still construct Markov chains, but now these chains can move between different models. The fraction of samples that the chain spends in each model is proportional to the evidence for that model and, in the case of models that differ only in the total number of sources, the evidences give the relative probabilities for the unknown number of sources in the data.

The most widely used algorithm for fitting multiple models is **reversible jump Markov**

chain Monte Carlo (RJMCMC). RJMCMC generates a Markov chain such that at each step either an update within the model is proposed, or, with a certain probability, a jump to an alternative model is proposed. Usually the jumps are between models that differ by only one source if that is the type of model hierarchy being considered. When proposing a jump to a new model, with parameters $\boldsymbol{\theta}'$, the values of the parameters of that model must also be proposed. This is achieved by generating a set of random numbers $\mathbf{u}$ from some distribution $q(\mathbf{u})$. In order to ensure reversibility we imagine that these random numbers are part of the parameters of the model, but because they are random we only need to generate them when they are used in a between-model jump. Similarly we may need some random variables $\mathbf{u}'$ to propose jumps back form the new model space to the original model parameters $\boldsymbol{\theta}$. The dimensionality of the joint space $(\boldsymbol{\theta}, \mathbf{u})$ must equal that of $(\boldsymbol{\theta}', \mathbf{u}')$ and there will be a deterministic, invertible mapping between the two. In the case of nested model, the reverse jump might just delete a set of parameters and so the dimensionality of $\mathbf{u}'$ is 0. However, if the particular source is deleted at random rather than, say, the lowest SNR source always being deleted, a random variable that selects which source to delete is required. The generalisation of the acceptance probability for RJMCMC is

$$\alpha = \min\left(1, \frac{p(\boldsymbol{\theta}'|\mathbf{x})q(\boldsymbol{\theta}')}{p(\boldsymbol{\theta}|\mathbf{x})q(\boldsymbol{\theta})}\left|\frac{\partial(\boldsymbol{\theta}', \mathbf{u}')}{\partial(\boldsymbol{\theta}, \mathbf{u})}\right|\right)$$

where the last term is the Jacobian for the transformation between the two sets of variables.

Example: mixture of Gaussians Suppose that model M_1 is a single Gaussian with mean θ_1 and unit variance and model M_2 is a mixture of two Gaussians with means θ_1' and θ_2' and both of unit variance. We have random variables $\mathbf{u} = (u_1, u_2)$ with $u_1 \sim N(0, \sigma_0^2)$ and $u_2 \sim U[0, 1]$ in the M_1 model space and $\mathbf{u}' = u_1' \sim U[0, 1]$ in the M_2 model space. The random variable u_1 gives the value of the mean of the new Gaussian to be added, while u_1' selects which Gaussian to delete in the reverse step. The second random variable u_2 ensures the dimensionality is consistent. We can define the mapping between the parameter spaces via

$$\begin{aligned}
\theta_1' &= \begin{cases} \theta_1 & \text{if } u_2 < 0.5 \\ u_1 \text{if } u_2 \geq 0.5 \end{cases} \\
\theta_2' &= \begin{cases} u_1 & \text{if } u_2 < 0.5 \\ \theta_1 \text{if } u_2 \geq 0.5 \end{cases} \\
u_1' &= u_2.
\end{aligned}$$

and the reverse mapping

$$\begin{aligned}
\theta_1 &= \begin{cases} \theta_1' & \text{if } u_2 < 0.5 \\ \theta_2' \text{if } u_1' \geq 0.5 \end{cases} \\
u_1 &= \begin{cases} \theta_2' & \text{if } u_2 < 0.5 \\ \theta_1' \text{if } u_1' \geq 0.5 \end{cases} \\
u_2 &= u_1'.
\end{aligned}$$

The Jacobian for this transformation is 1 and so the acceptance probability is just

$$\alpha = \min\left(1, \frac{p(\boldsymbol{\theta}'|\mathbf{x})q(\boldsymbol{\theta}')}{p(\boldsymbol{\theta}|\mathbf{x})q(\boldsymbol{\theta})}\right).$$

5.6 Evidence computation

As described earlier, the Bayesian evidence is required for model comparison and Bayesian model selection, but it is difficult to compute accurately using standard MCMC methods. **Nested sampling** (Skilling 2004) was developed as an alternative approach, specifically tuned for evidence computation. It calculates the evidence by transforming the multi–dimensional evidence integral into a one–dimensional integral that is easy to evaluate numerically. This is accomplished by defining the prior volume X as $dX = \pi(\mathbf{\Theta})d^D\mathbf{\Theta}$, so that

$$X(\lambda) = \int_{\mathcal{L}(\mathbf{\Theta})>\lambda} \pi(\mathbf{\Theta})d^N\mathbf{\Theta}, \tag{86}$$

where the integral extends over the region(s) of parameter space contained within the iso-likelihood contour $\mathcal{L}(\mathbf{\Theta}) = \lambda$. The evidence integral, Eq. (**??**), can then be written as

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X)dX, \tag{87}$$

where $\mathcal{L}(X)$, the inverse of Eq. (86), is a monotonically decreasing function of X. Thus, if one can evaluate the likelihoods $\mathcal{L}_i = \mathcal{L}(X_i)$, where X_i is a sequence of decreasing values,

$$0 < X_M < \cdots < X_2 < X_1 < X_0 = 1, \tag{88}$$

as shown schematically in Fig. 21, the evidence can be approximated numerically using standard quadrature methods as a weighted sum

$$\mathcal{Z} = \sum_{i=1}^{M} \mathcal{L}_i w_i, \tag{89}$$

where the weights w_i for the simple trapezium rule are given by $w_i = \frac{1}{2}(X_{i-1} - X_{i+1})$. An example of a posterior in two dimensions and its associated function $\mathcal{L}(X)$ is shown in Fig. 21.

5.6.1 Evidence Evaluation

The summation in Eq. (89) is performed as follows. The iteration counter is first set to $i = 0$ and N 'active' (or 'live') samples are drawn from the full prior $\pi(\mathbf{\Theta})$, so the initial prior volume is $X_0 = 1$. The samples are then sorted in order of their likelihood and the smallest (with likelihood $\mathcal{L}_0$) is removed from the active set (hence becoming 'inactive') and replaced by a point drawn from the prior subject to the constraint that the point has a likelihood $\mathcal{L} > \mathcal{L}_0$. The corresponding prior volume contained within this iso-likelihood contour will be a random variable given by $X_1 = t_1 X_0$, where t_1 follows the distribution $\mathbb{P}(t) = Nt^{N-1}$ (i.e., the probability distribution for the largest of N samples drawn uniformly from the interval $[0, 1]$). At each subsequent iteration i, the removal of the lowest likelihood point $\mathcal{L}_i$ in the active set, the drawing of a replacement with $\mathcal{L} > \mathcal{L}_i$ and the reduction of the corresponding prior volume $X_i = t_i X_{i-1}$ are repeated, until the entire prior volume has been traversed. The algorithm thus travels through nested shells of likelihood as the prior volume is reduced. The mean and standard deviation of $\log t$, which dominates the geometrical exploration, are:

$$E[\log t] = -1/N, \quad \sigma[\log t] = 1/N. \tag{90}$$

Since each value of $\log t$ is independent, after i iterations the prior volume will shrink down such that $\log X_i \approx -(i \pm \sqrt{i})/N$. Thus, one takes $X_i = \exp(-i/N)$.

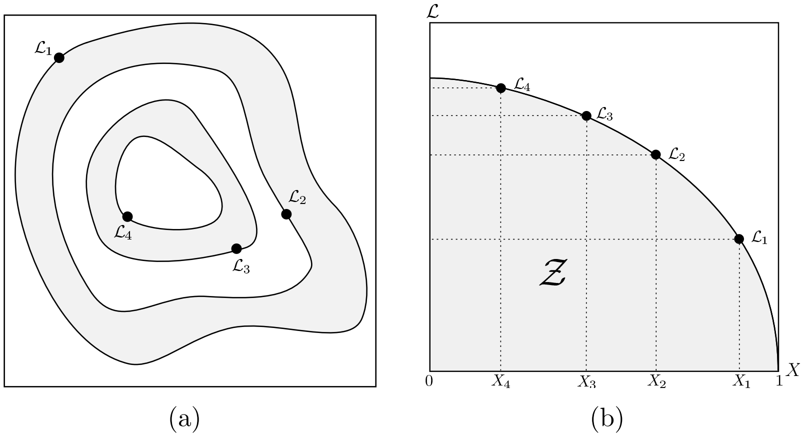


Figure 21: Cartoon illustrating (a) the posterior of a two dimensional problem; and (b) the transformed $\mathcal{L}(X)$ function where the prior volumes X_i are associated with each likelihood $\mathcal{L}_i$.

5.6.2 Stopping Criterion

The nested sampling algorithm should be terminated on determining the evidence to some specified precision. One way would be to proceed until the evidence estimated at each replacement changes by less than a specified tolerance. This could, however, underestimate the evidence in (for example) cases where the posterior contains any narrow peaks close to its maximum. Skilling provides an adequate and robust condition by determining an upper limit on the evidence that can be determined from the remaining set of current active points. By selecting the maximum–likelihood $\mathcal{L}_{\max}$ in the set of active points, one can safely assume that the largest evidence contribution that can be made by the remaining portion of the posterior is $\Delta\mathcal{Z}_{\rm i} = \mathcal{L}_{\max}X_{\rm i}$, i.e. the product of the remaining prior volume and maximum likelihood value. We choose to stop when this quantity would no longer change the final evidence estimate by some user–defined value (we use 0.5 in log–evidence).

5.6.3 Posterior Inferences

Once the evidence $\mathcal{Z}$ is found, posterior inferences can be easily generated using the final live points and the full sequence of discarded points from the nested sampling process, i.e., the points with the lowest likelihood value at each iteration i of the algorithm. Each such point is simply assigned the probability weight

$$p_i = \frac{\mathcal{L}_i w_i}{\mathcal{Z}}. \tag{91}$$

These samples can then be used to calculate inferences of posterior parameters such as means, standard deviations, covariances and so on, or to construct marginalised posterior distributions.

5.6.4 MultiNest Algorithm

The most challenging task in implementing the nested sampling algorithm is drawing samples from the prior within the hard constraint $\mathcal{L} > \mathcal{L}_i$ at each iteration i. Employing a

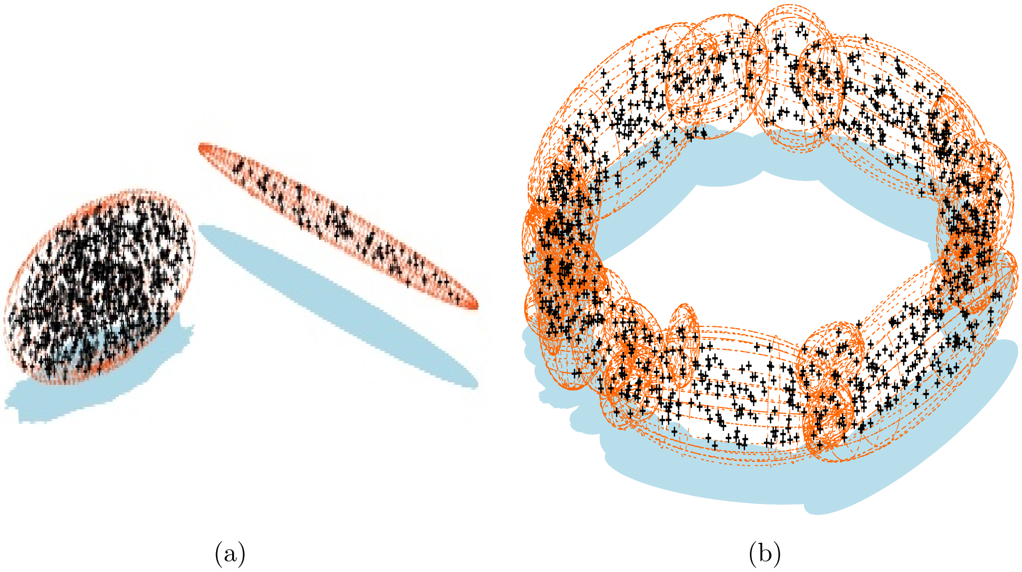


Figure 22: Illustrations of the ellipsoidal decompositions performed by MultiNest. The points given as input are overlaid on the resulting ellipsoids. 1000 points were sampled uniformly from: (a) two non-intersecting ellipsoids; and (b) a torus.

naive approach that draws blindly from the prior would result in a steady decrease in the acceptance rate of new samples with decreasing prior volume (and increasing likelihood). The MultiNest algorithm tackles this problem through an ellipsoidal rejection sampling scheme by enclosing the live point set within a set of (possibly overlapping) ellipsoids and a new point is then drawn uniformly from the region enclosed by these ellipsoids. The number of points in an individual ellipsoid and the total number of ellipsoids is decided by an 'expectation–maximization' algorithm so that the total sampling volume, which is equal to the sum of volumes of the ellipsoids, is minimized. This allows maximum flexibility and efficiency by breaking up a mode resembling a Gaussian into a relatively small number of ellipsoids, and if the posterior mode possesses a pronounced curving degeneracy so that it more closely resembles a (multi–dimensional) 'banana' then it is broken into a relatively large number of small 'overlapping' ellipsoids (see Fig. 22).

The ellipsoidal decomposition scheme described above also provides a mechanism for mode identification. By forming chains of overlapping ellipsoids (enclosing the live points), the algorithm can identify distinct modes with distinct ellipsoidal chains, e.g., in Fig. 22 panel (a) the algorithm identifies two distinct modes while in panel (b) the algorithm identifies only one mode as all the ellipsoids are linked with each other because of the overlap between them. Once distinct modes have been identified, they are evolved independently.

Another feature of the MultiNest algorithm is the evaluation of the global as well as the 'local' evidence values associated with each mode. These evidence values can be used in calculating the probability that an identified 'local' peak in the posterior corresponds to a real object.

There are many other nested sampling algorithms around today, including Polychord, which obtains samples from within the iso-likelihood surface through slice sampling, cpnest and dynesty. The latter two samplers form part of the Bilby parameter estimation software suite for LIGO.

6 Stochastic processes and sensitivity curves

In both frequentist and Bayesian approaches to statistical analysis, the likelihood plays a key role. This is the probability distribution from which the observed data has been drawn. In a gravitational wave context, we are typically concerned with analysing data from a noisy detector. The output from the detector, or detectors, is one or more real time series of measurements, $s_i(t)$. These measurements are a combination (usually assumed to be linear) of a signal part, $h_i(t)$, and a noise part, $n_i(t)$. The signal part is deterministic, depending only on the (unknown) parameters of the system, while the noise part is random. The likelihood is therefore a statement about the probability distribution from which the noise is drawn. The usual assumption is that the noise is generated by a **stationary**, **Gaussian random process**. In this section we will first define what this means, and discuss various approaches that are commonly used to summarise the noise properties and represent sensitivities to sources of different types.

6.1 Properties of random processes

A random process is a random sequence (often infinite in length) of values. Future values are not uniquely determined by current values, but by probability distributions that may be conditional on past values of the sequence. The observed random sequence is assumed to be drawn from *an ensemble of random processes* characterised by probability distributions

$$p_N(n_N, t_N; n_{N-1}, t_{N-1}; \ldots; n_2, t_2; n_1; t_1)\mathrm{d}n_N \mathrm{d}n_{N-1} \ldots \mathrm{d}n_2 \mathrm{d}n_1.$$

The probability distribution could be anything, but it is usual to make some simplifying assumptions, which are well motivated by observed random processes, to make computations plausible. The most commonly made assumptions are that the random process is **stationary**, **Gaussian** and **ergodic**.

A **stationary** random process is one for which the joint probability distributions for finite sets of samples depend only on time differences, not absolute time. In other words

$$p_N(n_N, t_N + \tau; \ldots; n_2, t_2 + \tau; n_1; t_1 + \tau) = p_N(n_N, t_N; \ldots; n_2, t_2; n_1; t_1) \; \forall \, \tau.$$

A random process is **Gaussian** if and only if all of its absolute probability distributions are Gaussian. In other words, for any set of N times, $\{t_1, \ldots, t_N\}$, we have

$$p_N(n_N, t_N; \ldots n_1; t_1) = A \exp\left[-\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N} \alpha_{jk}(n_j - \bar{n}_j)(n_k - \bar{n}_k)\right].$$

A ensemble of random process is **ergodic** if for any process, $n(t)$, drawn from the ensemble, the new ensemble defined by $\{n(t{+}KT) : \; k \in \mathbb{Z}\}$ has the same probability distributions.

To understand random processes, we are interested in both their mean values and the size of random fluctuations about the mean. We assume in the following (without loss of generality) that the mean of the random process is zero. Fluctuations about the mean can be characterised by the noise power (or variance), over a certain time interval $-T/2 < t < T/2$

$$\int_{-T/2}^{T/2} |n(t)|^2 \mathrm{d}t.$$

This quantity increase with time, linearly for stationary random processes. Therefore, it is more useful to work with the average value, referred to as the **mean power** or **mean square fluctuations**

$$P_n = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} |n(t)|^2 \mathrm{d}t.$$

It is useful to analyse quantities in the Fourier domain, and so we define

$$n_T(t) = n(t)\mathbb{I}\left[|t| < T/2\right],$$

which is just the full series truncated to the interval of interest. This notation allows us to use Parseval's Theorem

$$\int_{-T/2}^{T/2} [n(t)]^2 \mathrm{d}t = \int_{-\infty}^{\infty} [n_T(t)]^2 = \int_{-\infty}^{\infty} |\tilde{n}_T(f)|^2 \mathrm{d}f = 2\int_0^{\infty} |\tilde{n}_T(f)|^2 \mathrm{d}f$$

and we see that the mean square fluctuations are given by

$$P_n = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} [n(t)]^2 = \lim_{T\to\infty} \frac{2}{T}\int_0^{\infty} |\tilde{n}_T(f)|^2 \mathrm{d}f.$$

This motivates the definition of the spectral density, $S_n(f)$, via

$$S_n(f) = \lim_{T\to\infty} \frac{2}{T}\left|\int_{-T/2}^{T/2} n(t)\exp(2\pi i f t)\mathrm{d}t\right|^2.$$

This is the **one-sided spectral density**, which assumes that the time series is real and hence we only need to consider positive frequencies. The **two-sided spectral density**, which is is also defined for negative frequencies, is one half of this.

The spectral density represents the power in the process at a particular frequency since we have

$$P_n = \int_0^{\infty} S_n(f)\mathrm{d}f.$$

Suppose we are interested in the properties of the process in time intervals of length Δt, with corresponding **bandwidth** $\Delta f = 1/\Delta t$. The mean square fluctuations at frequency f in intervals of length Δt, and averaged over all intervals of that length, are

$$[\Delta n(\Delta t, f)]^2 \equiv \lim_{N\to\infty} \frac{2}{N}\sum_{n=-N/2}^{N/2}\left|\frac{1}{\Delta t}\int_{n\Delta t}^{(n+1)\Delta t} n(t)\exp(2\pi i f t)\mathrm{d}t\right|^2 = \frac{S_n(f)}{\Delta t} = S_n(f)\Delta f.$$

Hence we see that the *root mean square fluctuations at frequency f and measured over a time Δt* are just $\Delta n(\Delta t, f)_{\mathrm{rms}} = \sqrt{S_n(f)\Delta f}$. The spectral density can be interpreted in this way as the size of mean square fluctuations at the specified frequency.

A property of a random process that is closely linked to the spectral density is the **autocorrelation function**. This is defined in the standard way

$$C(\tau) = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} n(t)n(t+\tau)\mathrm{d}t.$$

For random processes that are ergodic (which implies they are also stationary), the averaging over time is equivalent to averaging over the ensemble

$$C(\tau) = \langle\, n(t)\, n(t+\tau)\,\rangle.$$

The auto-correlation function is the Fourier transform of the spectral density (the Wiener-Khinchin Theorem). A consequence of this result is that the expectation value of noise products can be written

$$\langle \tilde{n}^*(f)\tilde{n}(f')\rangle = S_n(f)\delta(f-f').$$

which is a statement that fluctuations of a stationary random process at different frequencies are uncorrelated with one another.

The spectral densities of a number of common noise processes are as follows

$$\begin{array}{ll} \textit{white noise spectrum} & S_n(f) = \text{const.} \\ \textit{flicker noise spectrum} & S_n(f) \propto 1/f \\ \textit{random walk spectrum} & S_n(f) \propto 1/f^2 \end{array} .$$

We conclude this section by noting that it is also possible to define a **cross-spectral density** between two separate random processes. This is defined via

$$S_{nm}(f) = \lim_{T\to\infty} \frac{2}{T}\left[\int_{-T/2}^{T/2} n(t)\exp(-2\pi i f t)\mathrm{d}t\right]\left[\int_{-T/2}^{T/2} m(t)\exp(2\pi i f t')\mathrm{d}t'\right]$$

and is the related via Fourier transform to the cross-correlation function of the two time series

$$C_{nm}(\tau) = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} n(t)m(t+\tau)\mathrm{d}t.$$

6.2 Sensitivity curves

For a Gaussian, stationary random process the spectral density conveys all the information about the statistical properties of the process. For gravitational wave detectors, it is therefore natural to plot the spectral density to characterise the detector sensitivity. But - how should sources be presented on the same diagram? There is no unique way to do this. Different types of source are best represented in different ways.

6.2.1 Burst signals

Burst signals are by definition compact in time duration, and usually also in frequency duration. It is rare that burst signals can be represented by parametric models, and so they are quite like random processes. We can characterise the burst by its frequency, f, duration, Δt, bandwidth, Δf, and its mean square amplitude, a proxy for the signal power

$$\bar{P}_h = \frac{1}{\Delta t}\int_0^{\Delta t} |h(t)|^2 \mathrm{d}t = h_c^2.$$

The square root of the mean square amplitude is called the **characteristic amplitude** of the burst. The power of the noise in the same bandwidth is $\Delta f S_n(f)$. The ratio of the power in the signal to the power in the noise is a measure of the detectability of the burst, relative to random fluctuations in the instrument. This ratio is the **signal-to-noise ratio** squared of the burst

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)^2 = \frac{\bar{P}_h}{\Delta f S_h(f)} = \frac{h_c^2}{\Delta f S_h(f)}.$$

If the data is windowed and bandpassed in the vicinity of the burst, then we maximise the contribution of the burst to the data and the signal-to-noise ratio is the ratio of the root-mean-square (rms) signal contribution to the rms noise contribution. For a broad-band burst with $\Delta f \sim f$ we have

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)^2 = \frac{h_c^2}{f S_h(f)}.$$

This motivates representing the sensitivity of a detector to bursts by plotting the quantity $f S_h(f)$ instead of the power spectral density. The detectability of a burst source with characteristic strain h_c can then be assessed by the height of h_c^2 above the curve.

6.2.2 Continuous sources

If instead of a burst we had a monochromatic gravitational wave source

$$h(t) = h_0 \exp(2\pi i f_0 t)$$

then the signal power is constant over time

$$P_h = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} |h(t)|^2 \mathrm{d}t = \frac{1}{2} h_0^2.$$

This power is concentrated at f_0. When observing a finite time series of length T, we can resolve frequencies to a precision $\Delta f \sim 1/T$. The noise power in that bandwidth is $S_n(f)/T$, which motivates representing the detector sensitivity curve by plotting

$$\sqrt{S_n(f)/T} \quad \text{or} \quad \rho_{\text{thresh}} \sqrt{S_n(f)/T}$$

where ρ_{thresh} is the threshold signal-to–noise ratio needed for detection. This is called the **strain spectral density**. The advantage of rep[resenting sensitivity in this way is that the detectability of a source can be directly assessed by seeing if the source amplitude h_0 lies above or below the curve. The height above the curve is a direct estimate of the signal to noise ratio of the source. The disadvantage of this way of representing sensitivity is that it varies with the length of observation, so this must be specified. In the case of LIGO, this is not a problem, as the detectors take periodic breaks from observation. After each observing run, the length of observation is known and so the strain spectral density can be evaluated for each observing run after the fact, and used to represent the results.

An example of a strain spectral density curve is given in Fig. 23.

Finally, we note that rescaling the sensitivity according to the detection threshold is not the only type of rescaled spectral density that is encountered in the literature. The amplitude of a gravitational wave signal in a gravitational wave detector depends on the orientation of the source relative to the detector plane. The same source placed at different sky locations and orientations will have different signal-to-noise ratios. To avoid having to specify which particular choices are being made, it is useful to produce a **sky-averaged sensitivity curve**. To assess detectability of a source, its amplitude should then be assessed

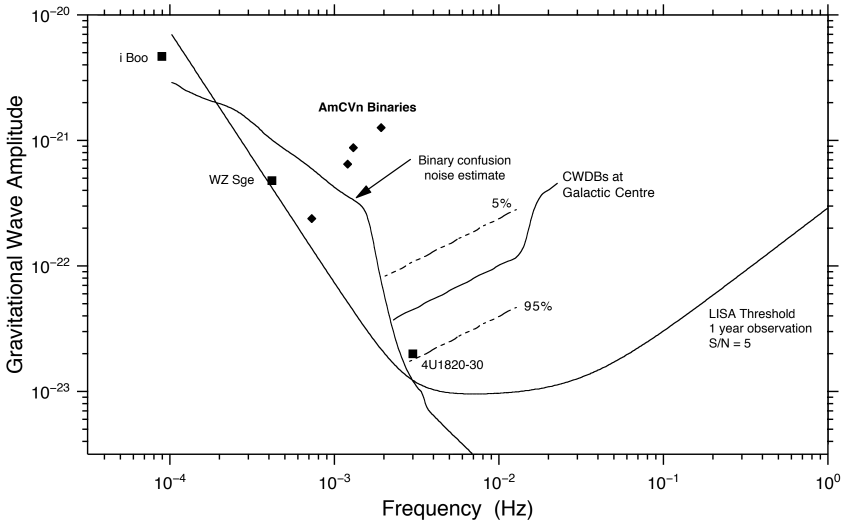


Figure 23: Strain spectral density curve for a 1 year observation with LISA and a detection threshold of $s/N = 5$. Reproduced from the LISA pre-phase A report.

for optimal orientation and sky location. The height of this optimal source above the sky averaged sensitivity is the average signal-to-noise ratio squared of a source of this type. For LIGO the sky averaged sensitivity is

$$\langle S_h(f) \rangle_{\mathrm{SA}}^{LIGO} \approx 5 S_h(f)$$

while for LISA we have

$$\langle S_h(f) \rangle_{\mathrm{SA}}^{LISA} \approx \frac{20}{3} S_h(f).$$

The difference arises because of the 60° opening angle of the LISA arms ($\sin^2 60 = 3/4$).

6.2.3 Inspiralling sources

Inspiraling sources have to be treated differently to continuous sources. This is because they emit a finite amount of power in each frequency band and hence the Fourier transform at each frequency is also finite. Therefore

$$\frac{1}{\sqrt{T}} \tilde{h}(f) \Rightarrow 0 \quad \text{as} \quad T \to \infty$$

and so the strain spectral density of an inspiraling source is zero averaged over all time. Band-passing and windowing the data can recover some signal-to-noise ratio, as in the burst source case, but we can do better than that using **filtering**.

A filtered time series is defined from a kernel $K(t - t')$ via convolution

$$w(t) = \int_{-\infty}^{\infty} K(t - t') s(t') \mathrm{d}t'.$$

In the previous cases we considered signal-to-noise ratio as the ratio of the rms power in the presence of a signal to the rms power in the noise. We use an analogous definition for filtered data, but now compare the amplitude of the filter output due to the signal only, to the rms amplitude of the filtered data in the presence of noise only

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)(t) = \frac{\int_{-\infty}^{\infty} K(t-t')h(t')\mathrm{d}t'}{\sqrt{\left\langle \left| \int_{-\infty}^{\infty} K(t-t')n(t')\mathrm{d}t' \right|^2 \right\rangle}}.$$

The rms output of the filter is the signal amplitude, S, to within a fractional error of N/S, which is the reciprocal of the signal-to-noise ratio.

The choice of the kernel is arbitrary, but it makes sense to choose the kernel that is "best" in some sense. The best kernel is the one that maximises the signal-to-noise ratio. This is most easily found by working in the Fourier domain. We use the Fourier transform definition

$$\tilde{x}(f) = \int_{-\infty}^{\infty} x(t)\exp\left[-2\pi i f t\right]\,\mathrm{d}t.$$

From the convolution theorem, the Fourier transform of the filter output is

$$\tilde{w}(f) = \tilde{K}(f)\tilde{h}(f)$$

where $\tilde{K}(f)$ and $\tilde{h}(f)$ are the Fourier transform of the kernel and waveform respectively. We have also

$$w(t) = \int_{-\infty}^{\infty} \tilde{x}(f)\exp\left[2\pi i f t\right]\,\mathrm{d}f \quad \Rightarrow \quad w(0) = \int_{-\infty}^{\infty} \tilde{x}(f)\,\mathrm{d}f.$$

Similarly

$$\begin{aligned}
N^2(0) &= \left\langle \left| \int_{-\infty}^{\infty} K(-t')n(t')\mathrm{d}t' \right|^2 \right\rangle = \left\langle \int_{-\infty}^{\infty} \tilde{K}(f)\tilde{n}(f)\mathrm{d}f \int_{-\infty}^{\infty} \tilde{K}^*(f')\tilde{n}^*(f')\mathrm{d}f' \right\rangle \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{K}(f)\tilde{K}^*(f')\left\langle \tilde{n}^*(f')\tilde{n}(f)\right\rangle \mathrm{d}f\mathrm{d}f' = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{K}(f)\tilde{K}^*(f')\delta(f-f')S_h(f)\mathrm{d}f\mathrm{d}f' \\
&= \int |\tilde{K}(f')|^2 S_h(f')\mathrm{d}f'. \qquad (92)
\end{aligned}$$

We deduce that the signal-to-noise ratio at zero lag is

$$\frac{\mathrm{S}}{\mathrm{N}} = \frac{\int \tilde{K}(f)\tilde{h}(f)\mathrm{d}f}{\sqrt{\int |\tilde{K}(f')|^2 S_h(f')\mathrm{d}f'}}$$

which can also be written as

$$\frac{\mathrm{S}}{\mathrm{N}} = \frac{(S_h K|h)}{\sqrt{(S_h K|S_h K)}}$$

by introducing the noise-weighted inner product

$$(\mathbf{h}_1|\mathbf{h}_2) = 2\int_0^{\infty} \frac{\tilde{\mathbf{h}_1}(f)\tilde{\mathbf{h}_2}^*(f) + \tilde{\mathbf{h}_1}^*(f)\tilde{\mathbf{h}_2}(f)}{S_h(f)}\,\mathrm{d}f.$$

This is of the form $\hat{\mathbf{e}}\cdot\mathbf{b}$, for a unit vector $\hat{\mathbf{e}}$ to be found. The inner product of two vectors of fixed length is maximised when they are parallel, i.e., $\hat{\mathbf{e}} \propto \mathbf{b}$. We therefore deduce that the choice which maximises the inner product is

$$S_h(f)\tilde{K}(f) \propto \tilde{h}(f) \quad \Rightarrow \quad \tilde{K}(f) \propto \frac{\tilde{h}(f)}{S_h(f)}.$$

This is the **Weiner optimal filter**. In the frequency domain the optimal filter is equal to the signal, weighted by the spectral density of the noise. A search using the optimal filter amounts to taking the inner product $(\mathbf{s}|\mathbf{h})$ of the data stream, $\mathbf{s}$, with a template of the signal, $\mathbf{h}$. This is **matched filtering**. In practice we don't know exactly what the signal is, but the parameters of the signal must be estimated from the data. In LIGO/Virgo this is done by computing the output of the optimal filter for a large number of source parameter choices which define a **template bank**.

The signal-to-noise ratio of the matched filtering search that uses the optimal filter is

$$\frac{\mathrm{S}}{\mathrm{N}}[\mathbf{h}] = \frac{(\mathbf{h}|\mathbf{h})}{\sqrt{\langle(\mathbf{h}|\mathbf{n})(\mathbf{h}|\mathbf{n})\rangle}} = (\mathbf{h}|\mathbf{h})^{1/2}$$

which follows from the fact that

$$\langle(\mathbf{h}_1|\mathbf{n})(\mathbf{h}_2|\mathbf{n})\rangle = (\mathbf{h}_1|\mathbf{h}_2). \tag{93}$$

This result is proved as follows

$$\begin{aligned}\langle(\mathbf{h}_1|\mathbf{n})(\mathbf{h}_2|\mathbf{n})\rangle &= \left\langle \int_{-\infty}^{\infty} \frac{\tilde{\mathbf{h}}_1(f)\tilde{\mathbf{n}}^*(f) + \tilde{\mathbf{h}}_1^*(f)\tilde{\mathbf{n}}(f)}{S_h(f)}\mathrm{d}f \int_{-\infty}^{\infty} \frac{\tilde{\mathbf{h}}_2(f')\tilde{\mathbf{n}}^*(f') + \tilde{\mathbf{h}}_2^*(f')\tilde{\mathbf{n}}(f')}{S_h(f')}\mathrm{d}f' \right\rangle \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\tilde{\mathbf{h}}_1(f)\tilde{\mathbf{h}}_2^*(f')\langle\tilde{\mathbf{n}}^*(f)\tilde{\mathbf{n}}(f')\rangle + \tilde{\mathbf{h}}_1^*(f)\tilde{\mathbf{h}}_2(f')\langle\tilde{\mathbf{n}}(f)\tilde{\mathbf{n}}^*(f')\rangle}{S_h(f)S_h(f')}\mathrm{d}f\mathrm{d}f' \\ &\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\tilde{\mathbf{h}}_1(f)\tilde{\mathbf{h}}_2(f')\langle\tilde{\mathbf{n}}^*(f)\tilde{\mathbf{n}}^*(f')\rangle + \tilde{\mathbf{h}}_1^*(f)\tilde{\mathbf{h}}_2^*(f')\langle\tilde{\mathbf{n}}(f)\tilde{\mathbf{n}}(f')\rangle}{S_h(f)S_h(f')}\mathrm{d}f\mathrm{d}f'.\end{aligned} \tag{94}$$

The terms on the final line vanish because $\langle\tilde{\mathbf{n}}(f)\tilde{\mathbf{n}}(f')\rangle = 0$, i.e., the size of fluctuations in the real and imaginary components of the noise are the same. The terms on the first line are simplified using $\langle\tilde{\mathbf{n}}^*(f)\tilde{\mathbf{n}}(f')\rangle = S_h(f)\delta(f-f')$

$$\begin{aligned}\langle(\mathbf{h}_1|\mathbf{n})(\mathbf{h}_2|\mathbf{n})\rangle &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{[\tilde{\mathbf{h}}_1(f)\tilde{\mathbf{h}}_2^*(f') + \tilde{\mathbf{h}}_1^*(f)\tilde{\mathbf{h}}_2(f')]\delta(f-f')}{S_h(f')}\mathrm{d}f\mathrm{d}f' \\ &= \int_{-\infty}^{\infty} \frac{[\tilde{\mathbf{h}}_1(f)\tilde{\mathbf{h}}_2^*(f) + \tilde{\mathbf{h}}_1^*(f)\tilde{\mathbf{h}}_2(f)]}{S_h(f)}\mathrm{d}f,\end{aligned} \tag{95}$$

giving the result stated above.

The matched filtering signal-to-noise ratio simplifies to

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)^2 = 4\int_0^{\infty} \frac{|\tilde{h}(f)|^2}{S_h(f)}\mathrm{d}f$$

which can also be written as

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)^2 = 4\int_0^{\infty} \frac{f|\tilde{h}(f)|^2}{S_h(f)}\mathrm{d}\ln f = 4\int_0^{\infty} \frac{f^2|\tilde{h}(f)|^2}{fS_h(f)}\mathrm{d}\ln f.$$

Plotting $S_h(f)$ and $f|\tilde{h}(f)|^2$ on a logarithmic frequency plot, the integral of the ratio of the two curves "by eye" gives an estimate of the signal-to-noise ratio squared.

For a source that has amplitude h_0 at frequency f, at which point the frequency derivative is $\dot{f}$, then the stationary phase approximation gives us the scaling

$$\tilde{h}(f) \sim \frac{h_0}{\sqrt{\dot{f}}}.$$

The analogy with the broad-band burst case described above motivates defining a characteristic strain, h_c, such that the signal-to-noise ratio squared is $h_c^2/(fS_h(f))$. The appropriate definition is

$$h_c = h_0\sqrt{\frac{2f^2}{\mathrm{d}f/\mathrm{d}t}} \sim f\tilde{h}(f).$$

Note also that since $\sqrt{\dot{E}} \sim \dddot{I}$ and $h \sim \ddot{I}/D$, we have $\sqrt{\dot{E}} \sim Dfh$ and hence

$$h_c \sim \frac{1}{D}\sqrt{\frac{\dot{E}}{\dot{f}}}$$

and this is an equality for monochromatic signals.

The characteristic strain is a measure of the signal-to-noise ratio accumulated while the frequency sweeps through a bandwidth equal to frequency. If we plot as a sensitivity curve the rms noise in a bandwidth equal to frequency, which is

$$h_n(f) \equiv \sqrt{f\langle S_h(f)\rangle_{\mathrm{SA}}\rangle}$$

then the signal-to-noise ratio accumulated as the inspiral proceeds from f to $2f$ is

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)^2_{f\to 2f} = \left[\frac{h_c(f)}{h_n(f)}\right]^2.$$

Therefore, plotting characteristic strain on the same plot gives a quick way to see how the signal-to-noise ratio of an inspiraling source builds up over the evolution. Note that plotting the characteristic strain only makes sense if the detector sensitivity is represented by $fS_h(f)$. If the detector sensitivity is represented by $S_h(f)$ then the quantity $h_c/\sqrt{f}$ should be used to represent the signal.

In the definition of characteristic strain, $h_c = h_0\sqrt{2f^2/\dot{f}}$, the term inside the square root is the number of cycles the inspiral spends in the vicinity of the frequency f. Papers that discuss matched filtering often include the statement that the signal to noise ratio is enhanced by the number of cycles spent in the vicinity of a certain frequency. This is what they are referring to.

In Fig. 24 we give an example of a plot of the characteristic strain, reproduced from Finn and Thorne (2000). The figure shows the characteristic strain of various extreme-mass-ratio inspiral sources detectable by LISA.

6.2.4 Stochastic backgrounds

Stochastic backgrounds are characterised by a spectral density, so it is natural to compute the power spectral density and plot it on the same axes as the detector PSD. However, there are two caveats. Firstly, the “power” we have been talking about so far has not been

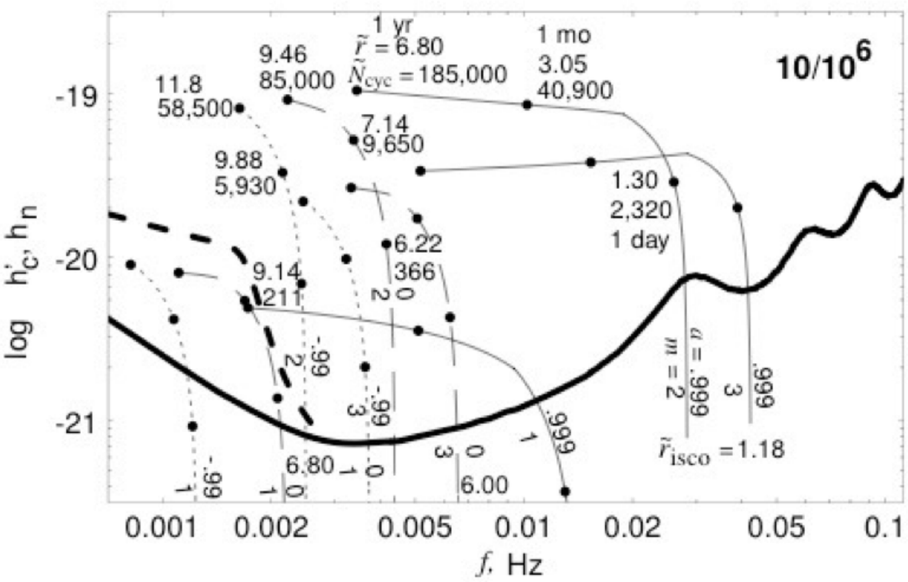


Figure 24: Characteristic strain for a number of typical extreme-mass-ratio inspiral sources observed by the classic (5 km arm length) LISA interferometer. All inspirals are circular, with $10^6 M_\odot$ central black holes and observed at a distance of 1Gpc. Curves are labelled by the spin of the central black hole, a, and the mass of the inspiraling object, m. Points on the curve correspond to 1 year, 1 month and 1 day prior to merger. The numbers above the points are the radius of the orbit (in units of M) at that time, and the number of gravitational wave cycles remaining until plunge. Reproduced from Finn and Thorne (2000).

a power in a physical sense since we have not specified any units for the time series (and indeed for GW strain this is dimensionless). When comparing to the noise power spectral density which is an energy density, it would be preferable to use something that represents a physical energy density if possible. Secondly, plotting two power spectral densities does not convey any information about their distinguishability. It would be preferable to represent a background in a way that conveys the detectability of the background at a glance.

The energy density carried by a gravitational wave is given by

$$\frac{\mathrm{d}E}{\mathrm{d}t\mathrm{d}A} \propto \dot{h}_+^2 + \dot{h}_\times^2.$$

Therefore, to obtain a physical energy density we should consider the time derivative of the strain. Differentiation with respect to time brings down a factor of frequency and so the energy spectral density is $f^2 S_h(f)$. Fluctuations of the energy spectral density in a bandwidth equal to frequency are then $f^3 S_h(f)$.

The energy density of an astrophysical or cosmological stochastic background, per logarithmic frequency interval, is often expressed as a fraction of the closure density of the Universe via

$$\Omega_{\mathrm{GW}} = \frac{8\pi G}{3H_0^2}\frac{\mathrm{d}E_{\mathrm{GW}}}{\mathrm{d}\ln f} \propto f^2 h_c^2(f).$$

The last equality defines the characteristic strain for a background, since, as argued above, a plane wave of frequency f and amplitude h_c carries an energy density $f h_c$. In the examples below we will show how to calculate the energy density for an astrophysical population of sources.

To represent backgrounds in a way that conveys their detectability directly, one can use *power-law sensitivity curves* (Thrane and Romano 2013). These are not uniquely defined, as they require some assumptions to be made about data analysis procedures and the threshold required for a detection using the defined procedure. However, given these assumptions, the procedure is as follows.

- For a given assumed power-law slope of a background, $\Omega_{\mathrm{GW}} \propto f^\beta$, compute the minimum amplitude, $A_{\min}(\beta)$, such that the background would be detectable by the defined procedure.

- Define the *power-law sensitivity curve*, $S_{\mathrm{pl}}(f)$, via

$$S_{\mathrm{pl}}(f) = \max\{A_{\min}(\beta) f^\beta \, : \, \beta \in [-\infty, \infty]\}.$$

The power-law sensitivity curve is the envelope of the minimal-detectable power-law backgrounds. It is a useful object to assess background detectability, since drawing a background of interest on the same figure gives an immediate indication of detectability. If the curve lies above the power-law sensitivity curve then it will be detectable (via the designated procedure) and otherwise it will not. An illustration of such a curve is given in Fig. 25.

6.3 Examples

We now estimate the leading order scaling of the quantities introduced above for some common astrophysical sources. Throughout we will make the usual choice of units to set $G = c = 1$.

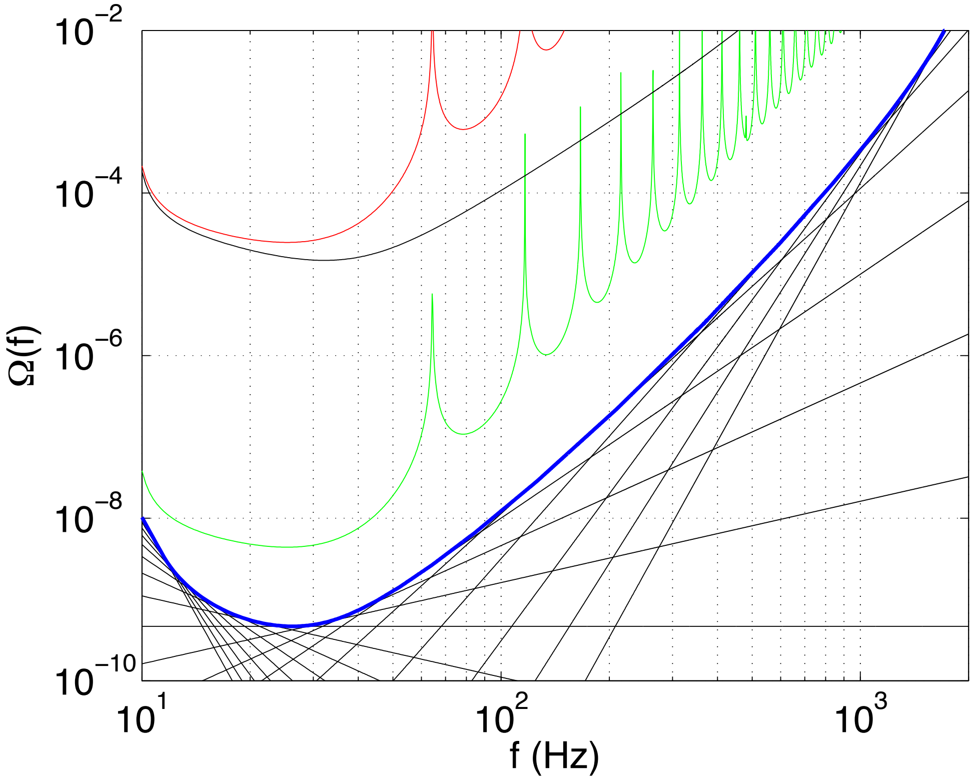


Figure 25: Power law sensitivity curve (in blue) for backgrounds detectable by ground-based interferometers, assuming the search is based on cross-correlation of the H1 and L1 detectors and the threshold for detection is a signal-to-noise ratio of 1. The solid black lines show the set of minimally-detectable power-laws that are used to generate the power-law sensitivity curve. The other curves show the detector strain spectral density instantaneously (red) and for a one year observation (green). Reproduced from Thrane and Romano (2013).

6.3.1 Single inspiraling compact binary

We consider first the case of compact binary coalescence. We assume that we have a circular binary with component masses M_1 and M_2 and separation r. We work in the Newtonian regime where the binary components are on Keplerian orbits. We denote the total mass, M, and reduced mass, μ, by

$$M = M_1 + M_2, \qquad \mu = \frac{M_1 M_2}{M_1 + M2}.$$

In the Newtonian two-body problem, the two objects each orbit around the centre of mass of the system. The two objects are distances r_1 and r_2 from the centre of mass respectively, where

$$r_1 M_1 = r_2 M_2 = \mu r.$$

The motion is also equivalent to that of a single body of mass μ orbiting in a fixed Newtonian potential of an object with mass M at a distance r. The orbital frequency is given by Kepler's laws

$$\omega^2 = \left(\frac{2\pi}{T}\right)^2 = (2\pi f)^2 = \frac{M}{r^3}.$$

To estimate the scaling of the gravitational wave emission we need to estimate the quadrupole moment of the binary. This can be estimated from

$$I \sim \mu r^2 \cos 2\omega t \sim \frac{M_1 M_2}{(M_1 + M_2)^{\frac{1}{3}}} \omega^{-\frac{4}{3}}.$$

At leading order, the gravitational wave strain scales like the second time derivative of the quadrupole moment divided by the distance to the source

$$h \sim \frac{\ddot{I}}{D} \sim \frac{1}{D} \frac{M_1 M_2}{(M_1 + M_2)^{\frac{1}{3}}} \omega^{\frac{2}{3}}.$$

The rate of energy loss scales like the third time derivative of I squared and so this has the scaling

$$\dot{E} \sim -\dddot{I}^2 \sim -\mu^2 M^{\frac{4}{3}} \omega^{\frac{10}{3}}.$$

Finally, we need to know how the energy relates to the orbital separation or equivalently the orbital frequency. In the Newtonian limit this follows from

$$E = -\frac{M\mu}{2r} = -\frac{\mu (M\omega)^{\frac{2}{3}}}{2}$$

from which we deduce

$$\dot{E} \sim -\mu M^{\frac{2}{3}} \omega^{-\frac{1}{3}} \dot{\omega}. \tag{96}$$

Combining this with expression (6.3.1) we obtain

$$\dot{\omega} \sim \mu M^{\frac{2}{3}} \omega^{\frac{11}{3}} = \frac{M_1 M_2}{(M_1 + M_2)^{\frac{1}{3}}} \omega^{\frac{11}{3}} = M_c^{\frac{5}{3}} \omega^{\frac{11}{3}}$$

where we have introduced the chirp mass

$$M_c = \frac{M_1^{\frac{3}{5}} M_2^{\frac{3}{5}}}{(M_1 + M_2)^{\frac{1}{5}}}.$$

We can now determine the scaling of the various quantities introduced in the previous section. From Eq. (6.2.3) and recalling $\omega = 2\pi f$, we obtain the Fourier domain amplitude

$$\tilde{h}(f) \sim \frac{h_0}{\sqrt{\dot{f}}} \sim \frac{1}{D}\frac{M_c^{\frac{5}{3}} f^{\frac{2}{3}}}{M_c^{\frac{5}{6}} f^{\frac{11}{6}}} = \frac{1}{D} M_c^{\frac{5}{6}} f^{-\frac{7}{6}}.$$

We can also deduce the characteristic strain

$$h_c(f) \sim \frac{1}{D} M_c^{\frac{5}{6}} f^{-\frac{1}{6}}.$$

6.3.2 Eccentric binaries

Eccentric binaries have gravitational wave emission at multiple harmonics of the orbital frequency (Peters and Matthews 1963). The flux of radiation at frequency nf, where n is the orbital frequency, is

$$\dot{E}_n = \frac{32}{5}\mu^2 M^{\frac{4}{3}} (2\pi f)^{\frac{10}{3}} g(n,e)$$

where $g(n,e)$ is given by

$$g(n,e) = \frac{n^4}{32}\left\{\left[J_{n-2}(ne) - 2eJ_{n-1}(ne) + \frac{2}{n}J_n(ne) + 2eJ_{n+1}(ne) - J_{n+2}(ne)\right]^2 \right.$$
$$\left. +(1-e^2)\left[J_{n-2}(ne) - 2J_n(ne) + J_{n+2}(ne)\right]^2 + \frac{4}{3n^2}\left[J_n(ne)\right]^2\right\} \qquad (97)$$

where $J_n(x)$ is the Bessel function of the first kind. The characteristic strain for an individual harmonic is therefore

$$h_{c,n}(f) = \frac{1}{\pi D}\sqrt{\frac{2\dot{E}_n(f/n)}{n\dot{f}(f/n)}} \sim M_c^{\frac{5}{6}} f^{-\frac{7}{6}} n^{\frac{2}{3}} \sqrt{g(n,e)}$$

where the argument (f/n) indicates that in order to get the contribution at frequency f from the n'th harmonic, it must be evaluated when the orbital frequency had the lower value of f/n.

It is normal to represent the contributions form individual waveform harmonics on a "waterfall plot". An example is shown in Figure 26 which is reproduced from Barack and Cutler (2004).

6.3.3 Stochastic backgrounds

The energy density in a gravitational wave background was defined in equation (6.2.4). If this background is generated by a population of individual sources, the total background can be estimated by integrating the contribution from each component in the background. The quantity of relevance is the total energy density in gravitational waves today, $\mathcal{E}_{\rm GW}$. If the sources are identical, have number density $n(z)$ and each generate a differential energy density $\mathrm{d}E/\mathrm{d}f$, then we have

$$\mathcal{E}_{\rm GW} = \int_0^\infty \rho_c c^2 \Omega_{\rm GW} \mathrm{d}\ln f = \int_0^\infty \int_0^\infty N(z)\frac{1}{(1+z)}\frac{\mathrm{d}E}{\mathrm{d}f} f \frac{\mathrm{d}f}{f}\mathrm{d}z,$$

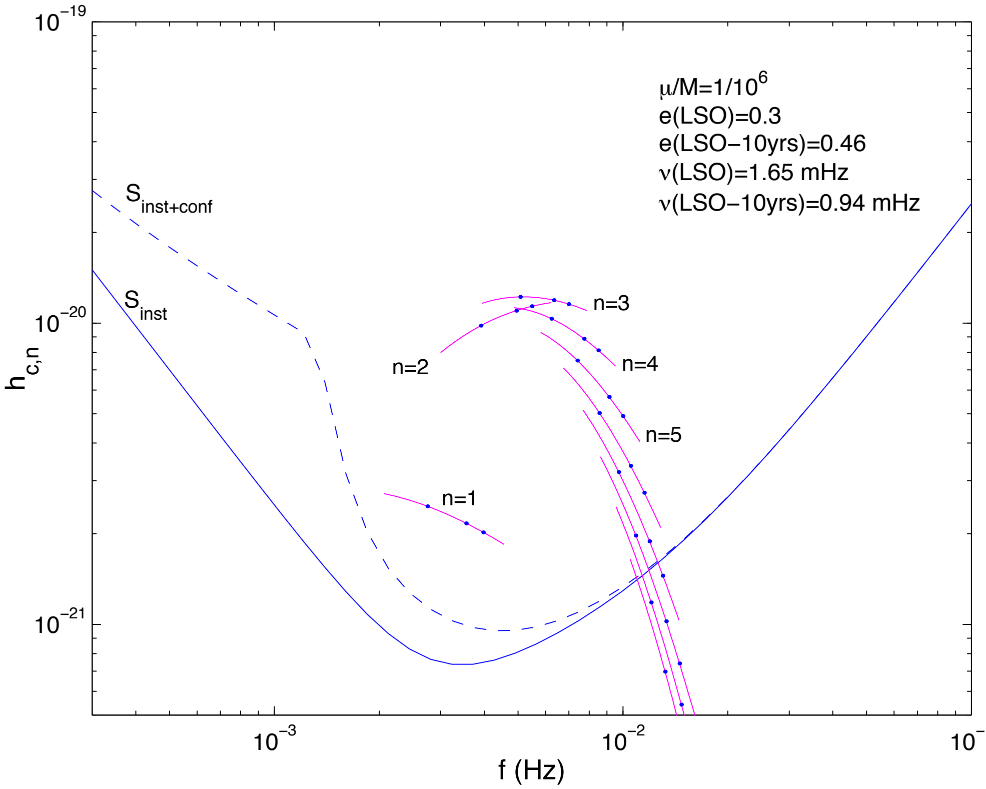


Figure 26: Characteristic strain of each harmonic in a the extreme-mass-ratio inspiral of a $1M_\odot$ black hole into a $10^6M_\odot$ black hole with eccentricity of 0.3 at plunge. Figure reproduced from Barack and Cutler (2004).

where the factor of $(1+z)$ accounts for the fact that the energy density today is redshifted relative to the energy density at emission. We deduce

$$\rho_c c^2 \Omega_{\rm GW} = \frac{\pi}{4}\frac{c^2}{G} f^2 h_c^2(f) = \int_0^\infty \frac{N(z)}{1+z}\left(f_r \frac{{\rm d}E}{{\rm d}f_r}\right)_{|f_r = f(1+z)} {\rm d}z \tag{98}$$

where the latter quantity is evaluated at the rest frame frequency, $f_r = (1+z)f$.

For a stochastic background generated by inspiraling binary sources, from Eq. (96), we have at leading order

$$f\frac{{\rm d}E}{{\rm d}f} \sim M_c^{\frac{5}{3}} f^{\frac{2}{3}}.$$

Plugging this into Eq. (98) we obtain

$$\Omega_{\rm GW}(f) \sim M_c^{\frac{5}{3}} f^{\frac{2}{3}} \int_0^\infty \frac{N(z)}{(1+z)^{\frac{1}{3}}} {\rm d}z. \tag{99}$$

We see that the energy spectral density of the background is

$$S_h(f) \sim \Omega_{\rm GW}(f)/f^3 \sim M_c^{\frac{5}{3}} f^{-\frac{7}{3}}$$

and the characteristic strain is

$$h_c(f) \sim \sqrt{\Omega_{\rm GW}(f)}/f \sim M_c^{\frac{5}{6}} f^{-\frac{2}{3}}.$$

In this case the characteristic strain scales like $f^{-2/3}$, while in the case of a single compact binary coalescence we had a scaling that was $f^{-1/6}$. This difference arises because the definition of characteristic strain relates to the signal to noise ratio that can be obtained in a search for the source of interest. For individual sources, we can perform matched filtering and enhance the signal to noise ratio coherently by the square root of the number of cycles (approximately $\sqrt{f}$, which explains the difference between $f^{-2/3}$ and $f^{-1/6}$). This is not possible for incoherent backgrounds where we can only predict the power at each frequency, not the phase.

7 Examples of frequentist statistics in gravitational wave astronomy

In this section we will describe some of the applications of frequentist statistical methods to gravitational wave detection. Fundamental to frequentist statistics is the likelihood. As described in the previous chapter, for gravitational wave detectors, we assume that the output of the detector, $s(t)$, is a linear combination of a signal, $h(t|\vec{\lambda})$, determined by a finite set of (unknown) parameters, $\vec{\lambda}$, and instrumental noise, $n(t)$. We assume in addition that the noise is Gaussian with a (usually known) power spectral density $S_h(f)$

$$s(t) = n(t) + h(t|\vec{\lambda}), \qquad \langle \tilde{n}^*(f)\tilde{n}(f')\rangle = S_h(f)\delta(f-f').$$

The signal is deterministic, but the noise is a random process. The likelihood, for parameters $\vec{\lambda}$, is therefore the probability that the observed noise realisation would take the value $n(t) = s(t) - h(t|\vec{\lambda})$, which can be seen to be

$$\mathcal{L}(s|\vec{\lambda}) = p(n(t) = s(t) - h(t|\vec{\lambda})) \propto \exp\left[-\frac{1}{2}(s - h(\vec{\lambda})|s - h(\vec{\lambda}))\right] \tag{100}$$

where the noise weighted overlap is as given in the last lecture

$$(a|b) = \int_{-\infty}^{\infty} \frac{\tilde{a}^*(f)\tilde{b}(f) + \tilde{a}(f)\tilde{b}^*(f)}{S_h(f)} \mathrm{d}f.$$

7.1 The Fisher Matrix

We introduced the Fisher Matrix in the discussion of the Cramer-Rao bound on the variance of an estimator, which, for a multivariate unbiased estimator, $\hat{\lambda}$, is given by

$$\mathrm{cov}(\hat{\lambda}_i, \hat{\lambda}_j) \geq [\mathbf{\Gamma}_\lambda]^{-1}_{ij}$$

where

$$(\mathbf{\Gamma}_\lambda)_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \lambda_i}\frac{\partial l}{\partial \lambda_j}\right].$$

In the above l denotes the log-likelihood. For the gravitational wave log-likelihood in Eq. (100), the derivative is

$$\frac{\partial l}{\partial \lambda_i} = \left(\frac{\partial h}{\partial \lambda_i}\middle| s - h(\vec{\lambda})\right) = \left(\frac{\partial h}{\partial \lambda_i}\middle| \mathbf{n}\right).$$

It therefore follows, from the result given in Eq. (93), that

$$(\mathbf{\Gamma}_\lambda)_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \lambda_i}\frac{\partial l}{\partial \lambda_j}\right] = \left\langle \left(\frac{\partial h}{\partial \lambda_i}\middle|\mathbf{n}\right)\left(\frac{\partial h}{\partial \lambda_i}\middle|\mathbf{n}\right)\right\rangle = \left(\frac{\partial h}{\partial \lambda_i}\middle|\frac{\partial h}{\partial \lambda_j}\right).$$

The Fisher Matrix gives a lower bound on the variance of any unbiased estimator of the parameters of the signal, and hence it provides a guide to how accurately the parameters can be measured. We know that the maximum likelihood estimator is asymptotically efficient, i.e., it achieves this Fisher Matrix bound, which is why it might be expected to provide a good guide to parameter measurement precision. However, asymptotic efficiency refers to making many repeated measurements of the same parameter, which we do not typically do in gravitational wave observations. But it can be seen that the Fisher Matrix provides a good guide to measurement precision even for a single observation, as follows. We suppose that the true parameters of the signal are given by $\vec{\lambda}_0$, and expand to leading order about those parameters

$$\vec{\lambda} = \vec{\lambda}_0 + \Delta\vec{\lambda}, \qquad h(t|\vec{\lambda}) = h(t|\vec{\lambda}) + \partial_i h(t|\vec{\lambda})\Delta\lambda^i$$

where ∂_i denotes the derivative with respect to λ_i and the last term employs Einstein summation convention. This approximation is known as the **linear signal approximation**. The likelihood can then be expanded as

$$\begin{aligned}
\mathcal{L}(s|\vec{\lambda}) &\propto \exp\left[-\frac{1}{2}(n - \partial_i h(t|\vec{\lambda})\Delta\lambda^i | n - \partial_j h(t|\vec{\lambda})\Delta\lambda^j)\right] \\
&= \exp\left\{-\frac{1}{2}\left[(n|n) - 2(n|\partial_i h(t|\vec{\lambda}))\Delta\lambda^i + (\partial_i h(t|\vec{\lambda})|\partial_j h(t|\vec{\lambda}))\Delta\lambda^i\Delta\lambda^j\right]\right\} \\
&= \exp\left[-\frac{1}{2}(n|n)\right] \exp\left[-\frac{1}{2}\left(\Delta\lambda^i - (\Gamma^{-1})_{ik}(n|\partial_k h(t|\vec{\lambda}))\right)\Gamma_{ij}\left(\Delta\lambda^j - (\Gamma^{-1})_{jl}(n|\partial_l h(t|\vec{\lambda}))\right)\right] \\
&\qquad \times \exp\left[-\frac{1}{2}(n|\partial_i h(t|\vec{\lambda}))(\Gamma^{-1})_{ij}(n|\partial_j h(t|\vec{\lambda}))\right]. \qquad (101)
\end{aligned}$$

The latter term is sub-dominant since it is $O(1)$ compare to the middle term which is of order of the signal amplitude, or SNR. The middle term is a Gaussian, centred at $\Delta\lambda^i = (\Gamma^{-1})_{ik}(n|\partial_k h(t|\vec{\lambda}))$, and with covariance matrix given by the Fisher Matrix. The latter therefore provides an estimate of the width of the likelihood distribution and hence can be used as a guide to the uncertainty. In addition, the maximum likelihood estimator

$$\widehat{\Delta\lambda}^i = (\Gamma^{-1})_{ik}(n|\partial_k h(t|\vec{\lambda}))$$

has mean and variance

$$\mathbb{E}\left(\widehat{\Delta\lambda}^i\right) = 0, \qquad \mathrm{cov}\left(\widehat{\Delta\lambda}^i, \widehat{\Delta\lambda}^j\right) = \Gamma^{-1}_{ij},$$

which again confirms the interpretation of the Fisher Matrix as the uncertainty in the parameter estimate. The fractional corrections to the Fisher Matrix estimate scale like the inverse of the signal-to-noise ratio and therefore the Fisher Matrix is a good approximation in the high signal-to-noise ratio limit.

The Fisher Matrix has been widely used in a gravitational wave context to assess the measurability of parameters using observations with present or future detectors. While the Fisher Matrix is only an approximation, it can be directly calculated by evaluating a small number of waveforms, rather than requiring samples to be obtained all over the waveform parameter space, and so it is much cheaper computationally. This makes it a good tool for Monte Carlo simulations over parameter space, to survey parameter estimation accuracies over a wide parameter range.

7.2 Matched filtering

In the previous chapter we introduced the idea of matched filtering, motivated by maximising the signal to noise ratio of a filtered data stream. The optimal filter has a frequency-domain kernel $\tilde{K}(f) \propto \tilde{h}(f)/S_h(f)$. The use of the output of the optimal filter as a test statistic for a search can also be motivated by the frequentist concepts that we encountered in previous chapters. Suppose that we write $\mathbf{h}(\lambda) = A\hat{\mathbf{h}}(\lambda)$, where $(\hat{\mathbf{h}}(\lambda)|\hat{\mathbf{h}}(\lambda)) = 1$, to separate out the amplitude of the gravitational wave source from the other parameters. The log-likelihood can be written

$$\begin{aligned} l(\lambda) &= -\frac{1}{2}(\mathbf{s} - A\hat{\mathbf{h}}(\lambda)|\mathbf{s} - A\hat{\mathbf{h}}(\lambda)) = -\frac{1}{2}\left[(\mathbf{s}|\mathbf{s}) - 2A(\mathbf{s}|\hat{\mathbf{h}}) + A^2\right] \\ &= -\frac{1}{2}\left[(\mathbf{s}|\mathbf{s}) + (A - (\mathbf{s}|\hat{\mathbf{h}}))^2 - (\mathbf{s}|\hat{\mathbf{h}})^2\right]. \end{aligned} \tag{102}$$

Fora given λ, this is maximized by the choice $A = (\mathbf{s}|\hat{\mathbf{h}})$, for which the log-likelihood $\propto (\mathbf{s}|\hat{\mathbf{h}})^2 - (\mathbf{s}|\mathbf{s})$. The maximum likelihood estimator for parameters other than the amplitude is thus given by the maximum of the optimal filter output over the parameter space. So, optimal filtering is just maximum likelihood estimation. To do this in practice, the optimal filter must be evaluated over the whole parameter space. In the analysis of gravitational wave data, from LIGO in particular, this is achieved using a **template bank**, which is a set of templates that cover the whole parameter space. The overlap of each template with the detector data is evaluated, and the maximum of those template overlaps is used as a test statistic to identify whether or not there is a signal in the data.

The question that we want to ask is "Is there a gravitational wave signal in the data?". Assuming that the parameters λ are fixed, this can formulated as a hypothesis test on the signal amplitude

$$H_0 : A = 0, \qquad \text{vs.} \qquad H_1 : A > 0.$$

From the Neyman-Pearson lemma the optimal statistic for testing the simple hypothesis $A = 0$ versus $A = A_1$ is the likelihood ratio, which is

$$\exp\left[A_1(\mathbf{s}|\hat{\mathbf{h}}(\lambda)) - \frac{1}{2}A_1^2\right].$$

This is large for large values of the optimal filter $(\mathbf{s}|\hat{\mathbf{h}}(\lambda))$ and so we deduce that the optimal filter is also the most powerful detection statistic. As the detection statistic does not depend on A_1, this test is uniformly most powerful for the composite hypothesis $A > 0$. In the more usual case that λ is unknown, although the maximum of the optimal filter statistic is still the maximum likelihood estimator, this is no longer a uniformly most powerful test, although it remains quite close to being so.

LIGO matched filtering searches typically use a large number of templates, distributed throughout the parameter space in a **template bank**. The matched filter output is evaluated for all of these templates, and the maximum filter output over the template bank is used as a detection statistic. Template banks are typically characterised by their **minimal match**, MM. This is defined as the *minimum* over all *possible signals* of the **maximum** overlap of that signal with one of the templates in the bank

$$\min_{\vec{\lambda}}\left[\max_{h_{\rm temp,i}:i=1,\ldots,N}(h(\vec{\lambda})|h_{\rm temp,i})\right] \gtrsim \text{MM}$$

where $\{h_{\rm temp,i} : i = 1, \ldots, N\}$ are the N templates in the template bank. The minimal match is the worst possible detection statistic that a randomly chosen signal could have. Setting this minimal match to some value close to 1 ensures that very few signals will be missed. A typical value of the minimal match used in practice would be 0.97. For a uniform distribution of sources in a Euclidean Universe, the fraction of sources that would be missed is $1 - 0.97^3 = 0.087$.

Template banks can be constructed analytically using the Fisher Matrix as a metric. This follows from expanding the overlap of two normalised templates, $\hat{h}(\vec{\lambda}) = h(\vec{\lambda})/\sqrt{(h(\vec{\lambda})|h(\vec{\lambda}))}$,

$$(\hat{h}(\vec{\lambda})|\hat{h}(\vec{\lambda}+\Delta\vec{\lambda})) = (\hat{h}(\vec{\lambda})|\hat{h}(\vec{\lambda})) + \left(\hat{h}(\vec{\lambda})\middle|\frac{\partial\hat{h}}{\partial\lambda_i}(\vec{\lambda})\right)\Delta\lambda^i + \frac{1}{2}\left(\hat{h}(\vec{\lambda})\middle|\frac{\partial^2\hat{h}}{\partial\lambda_i\partial\lambda_j}(\vec{\lambda})\right)\Delta\lambda^i\Delta\lambda^j + \cdots .$$

The first term is 1 because of the normalisation. The second term vanishes since

$$(\hat{h}(\vec{\lambda})|\hat{h}(\vec{\lambda})) = 1 \quad \Rightarrow \quad \frac{\partial}{\partial\lambda_i}(\hat{h}(\vec{\lambda})|\hat{h}(\vec{\lambda})) = 0 \quad \Rightarrow \quad \left(\hat{h}(\vec{\lambda})\middle|\frac{\partial\hat{h}}{\partial\lambda_i}(\vec{\lambda})\right) = 0.$$

The third term can be simplified using

$$\begin{aligned} \frac{\partial}{\partial\lambda_j}\left(\hat{h}(\vec{\lambda})\middle|\frac{\partial\hat{h}}{\partial\lambda_i}(\vec{\lambda})\right) = 0 \quad &\Rightarrow \quad \left(\frac{\partial\hat{h}}{\partial\lambda_i}(\vec{\lambda})\middle|\frac{\partial\hat{h}}{\partial\lambda_j}(\vec{\lambda})\right) + \left(\hat{h}(\vec{\lambda})\middle|\frac{\partial^2\hat{h}}{\partial\lambda_i\partial\lambda_j}(\vec{\lambda})\right) = 0 \\ &\Rightarrow \quad \left(\frac{\partial\hat{h}}{\partial\lambda_i}(\vec{\lambda})\middle|\frac{\partial\hat{h}}{\partial\lambda_j}(\vec{\lambda})\right) = -\left(\hat{h}(\vec{\lambda})\middle|\frac{\partial^2\hat{h}}{\partial\lambda_i\partial\lambda_j}(\vec{\lambda})\right). \end{aligned} \tag{103}$$

We deduce

$$(\hat{h}(\vec{\lambda})|\hat{h}(\vec{\lambda}+\Delta\vec{\lambda})) = 1 - \frac{1}{2}\Gamma_{ij}\Delta\lambda^i\Delta\lambda^j.$$

The Fisher Matrix (of normalised templates) thus provides a metric on parameter space, which can be used to place templates. This is only practical in low numbers of dimensions. In higher numbers of dimensions, it is easier to use **stochastic template banks**. A stochastic bank is constructed as follows

1. At step 1, choose the first template, $\hat{h}(\lambda_1)$, randomly from parameter space. Add it to the template bank, $\mathcal{T}$.

2. At step $i \geq 2$, set the counter to 1 and then repeat the following steps:

   (a) Draw a random set of parameter values, $\vec{\lambda}_i$, and evaluate the match, M, with the current template bank

   $$M = \left[\max_{h_{\rm temp}\in\mathcal{T}}(h(\vec{\lambda}_i)|h_{\rm temp})\right].$$

   (b) If $M < MM$, add $h(\vec{\lambda}_i)$ to the template bank and advance to step $i+1$. Otherwise, increment the counter. If the counter has reached $N_{\max}$, stop. Otherwise return to step (a).

7.3 LIGO searches

LIGO employs two different matched filtering algorithms to search for signals, Pycbc and Gstlal. They differ in various details, including how the template overlaps are computed. We will not discuss these in detail here, but refer the interested reader to relevant publications. For *gstlal* these are

- Cannon, K., Cariou, R., Chapman, A., et al. (2012), *Astrophys. J.* **748**, 136, doi: 10.1088/0004-637X/748/2/136.
- Privitera, S., Mohapatra, S. R. P., Ajith, P., et al. (2014), *Phys. Rev.* D **89**, 024003, doi: 10.1103/PhysRevD.89.024003
- Messick, C., Blackburn, K., Brady, P., et al. (2017), *Phys. Rev.* D **95**, 042001, doi: 10.1103/PhysRevD.95.042001
- Sachdev, S., Caudill, S., Fong, H., et al. (2019), arXiv:1901.08580
- Hanna, C., Caudill, S., Messick, C., et al. (2019), arXiv:1901.02227

For *pycbc* the relevant references are

- Nitz, A., Harry, I., Brown, D., et al. (2019), gwastro/pycbc: PyCBC Release v1.15.2, doi: 10.5281/zenodo.3596447
- Nitz, A. H., Dal Canton, T., Davis, D., & Reyes, S. (2018), *Phys. Rev.* D **98**, 024050, doi: 10.1103/PhysRevD.98.024050
- Usman, S. A., Nitz, A. H., Harry, I. W., et al. (2016), *Class. Quantum Grav.* **33**, 215004, doi: 10.1088/0264-9381/33/21/215004

Both searches adopt a traditional frequentist framework, in that the output of the pipeline is used as a detection statistic. If the detection statistic exceeds a threshold then the data is flagged as interesting, i.e., potentially containing a signal. The threshold is determined based on the behaviour of the search pipeline in the absence of any signals in the data. This background distribution is estimated using **time slides**. Both searches rely on consistency between triggers in two or more detectors. Any astrophysical gravitational wave signal must pass through both detectors within an interval of 10ms. If the data of one detector is time shifted relative to the other by more than this amount, then any coincident triggers in the two instruments must be due to instrumental noise only. By doing many different time shifts in this way, the background distribution can be estimated for much longer effective observation times.

In hypothesis testing, we discussed the notion of a significance or p-value. This makes sense if the size of the data set is fixed, but gravitational wave detectors are continuously taking data. Therefore it makes sense to quantify significance instead by a *false alarm rate* or FAR, which is the frequency at which triggers as extreme as the one observed, or more extreme, occur in the data. LIGO quotes FARs for all events that are distributed publicly.

We will now give an overview of a few techniques that are used in LIGO searches to improve their speed and efficiency.

7.3.1 Waveform consistency

The assumptions that lead to the optimal filter assume that the noise is stationary. This is approximately true for gravitational wave detectors, but they are also observed to have large glitches quite often. While the glitches do not match any of the templates well, there is often sufficient power in the glitch that they can trigger the detection statistic to exceed the threshold. To mitigate for this problem, LIGO searches use **waveform consistency** checks. These verify that after subtracting the best-fit template signal from the data, the resulting time series is consistent with being stationary Gaussian noise with the estimated PSD. If the template $\hat{h}$ coincides with the true signal, the quantity

$$\chi^2 = \sum_{k=1}^{N} \frac{|\hat{s}_k - \hat{h}_k|^2}{S_h(f_k)}$$

is the sum of squares of $N(0,1)$ distributed random variables, and hence follows a chi-squared distribution with N degrees of freedom. The mean of a χ^2_N random variable is N, so χ^2/N should be expected to be close to 1 if the template is a good match to the data, and much bigger otherwise. LIGO uses something called *effective SNR* as a detection statistic. This is defined as

$$\hat{\rho} = \frac{\rho}{(1+(\chi^2/N)^3)^{\frac{1}{6}}}.$$

For real signals, this is close to the true SNR, while for glitches it is much smaller. The effective SNR is used as the detection statistic by *pycbc*.

7.3.2 Marginalisation over phase and time

A template bank requires templates in all parameters, so it is useful to reduce the dimensionality of the parameter space whenever possible. This can be done straightforwardly for the *initial phase* and *time of coalescence*. For a monochromatic signal

$$h(t|A, f_0, t_c, \phi_0) = A\cos(2\pi f_0(t-t_c)+\phi_0) = A\cos(2\pi f_0(t-t_c))\cos\phi_0 - A\sin(2\pi f_0(t-t_c)\sin\phi_0$$

the matched filter overlap is

$$(s|h) = A\cos\phi_0 O_c - A\sin\phi_0 O_s, \qquad \text{where } O_c = (s|\cos(2\pi f_0(t-t_c))), \quad O_s = (s|\sin(2\pi f_0(t-t_c))).$$

Differentiating with respect to ϕ_0 and equating it to zero, we find that the value of ϕ_0 that maximises the overlap is

$$\tan\phi_0 = -\frac{O_s}{O_c} \qquad \Rightarrow \qquad \max_{\phi_0}(s|h)^2 = A^2(O_c^2 + O_s^2).$$

If this is used instead of the standard overlap, then the template bank automatically maximises over phase and this parameter direction does not need to be covered by templates.

To maximize over the unknown coalescence time we use

$$\tilde{h}(f|A, f_0, t_c, \phi_0) = \tilde{h}(f|A, f_0, 0, \phi_0)\exp(-2\pi i f t_c)$$

and observe that

$$(s|h(t|A, f_0, t_c, \phi_0)) = 2\Re\int_{-\infty}^{\infty} \frac{\tilde{s}^*(f)\tilde{h}(f|A, f_0, 0, \phi_0)}{S_h(f)}\exp(-2\pi i f t_c)\mathrm{d}f.$$

This is just the inverse Fourier transform of

$$\frac{\tilde{s}^*(f)\tilde{h}(f|A, f_0, 0, \phi_0)}{S_h(f)}.$$

Inverse Fourier transforms can be computed cheaply (in $n \log n$ time) using the fast Fourier transform. Therefore, the time of coalescence can be efficiently maximized over by computing the quantity above, taking its inverse fast Fourier transform, and then finding the maximum of the components of the resulting vector.

7.3.3 The F-statistic

The F-statistic is an extension of the above ideas to more of the extrinsic parameters of the signal. It is not used so much for LIGO, but has been used extensively in LISA data analysis work (see for example Cornish & Porter (2007), *Phys. Rev.* D**75**, 021301; *Class. Quantum Grav.* **24**, 5729). The idea is to write the signal as a sum of modes, such that the coefficients depend only on a (subset of) the extrinsic parameters, and then analytically maximise over those coefficients. For SMBH binaries in LISA the decomposition takes the form

$$h(t) = \sum_{i=1}^{4} a_i(\iota, \psi, D_L, \phi_c) A^i(t|M_c, \mu, t_c, \theta, \phi)$$

where

$$\begin{aligned}
a_1 &= \Lambda[(1+\cos^2\iota)\cos 2\psi \cos\phi_c - 2\cos\iota \sin 2\psi \sin\phi_c] \\
a_2 &= -\Lambda[(1+\cos^2\iota)\sin 2\psi \cos\phi_c + 2\cos\iota \cos 2\psi \sin\phi_c] \\
a_3 &= \Lambda[(1+\cos^2\iota)\cos 2\psi \sin\phi_c + 2\cos\iota \sin 2\psi \cos\phi_c] \\
a_4 &= -\Lambda[(1+\cos^2\iota)\sin 2\psi \sin\phi_c - 2\cos\iota \cos 2\psi \cos\phi_c] \\
A_1 &= M\eta x(t) D^+ \cos(\Phi) \\
A_2 &= M\eta x(t) D^\times \cos(\Phi) \\
A_3 &= M\eta x(t) D^+ \sin(\Phi) \\
A_4 &= M\eta x(t) D^\times \sin(\Phi).
\end{aligned} \tag{104}$$

Here the waveform parameters are inclination ι, polarization angle, ψ, luminosity distance, D_L, phase at coalescence, ϕ_c, chirp mass, M_c, reduced mass ratio, μ, time of coalescence, t_c, colatitude, θ, and azimuth, ϕ. We denote the waveform phase by $\Phi(t)$ and $x = (GM\omega/c^3)^{2/3}$, where ω is the orbital frequency and $M = m_1 + m_2$ is the total mass. The quantities D^+ and D^x are the two components of LISA's time-dependent response function.

Writing $N^i = (s|A^i)$, the matched filter overlap is

$$(s|h) = a_j N^j$$

and we want to maximise this subject to the constraint that the waveform is normalised which becomes

$$a_i M^{ij} a_j = 1,' \qquad \text{where } M^{ij} = (A^i|A^j).$$

This is a standard optimisation problem with solution

$$a_i = (M^{-1})_i j N^j = M_{ij} N^j.$$

The maximized value of the log-likelihood is the F-staistic

$$\mathcal{F} = \frac{1}{2} M_{ij} N^i N^j.$$

This can be used to automatically maximise over extrinsic parameters in a search, reducing the dimensionality of the parameter space to just that of the intrinsic parameters. Note that in the above we have taken the coefficients, a_i, to be independent of one another and unconstrained, while in practice they are correlated and take a potentially limited range of values because they all depend on the same set of four extrinsic parameters. Thus, we are finding the maximum over a space that is somewhat larger than the true space, and contains some unphysical values. If there is a signal in the data, then the maximization must nonetheless still give the right extrinsic parameter values (in the absence of noise).

7.3.4 Power spectral density estimation

The likelihood contains the spectral density of noise in the detector, which is usually not known precisely. LIGO searches (and parameter estimation codes) need to use a PSD that has been estimated from the data. This is accomplished by considering a number of other sections of data, distributed either side of the section of data that is of interested because it is believed to contain a signal. The power spectrum (i.e., the norm squared of the Fourier transform) is computed for each of the empty segments, $\sigma_i^2(f)$, and then these can be combined to give an estimate of the PSD in the segment of interest. The averaging can be done by taking the mean

$$\sigma_0^2(f) = \frac{1}{2N} \sum_{k=1}^{N} (s_k^2 + s_{-k}^2)$$

but in LIGO analyses it is more usual to use the median. The median is less susceptible to outliers in the data arising from non-stationary features in the noise.

7.4 Unmodelled searches

For burst sources matched filtering cannot be used, as it is not possible to build templates of potential signals. LIGO uses a number of different searches for unmodelled sources. Again, we won't describe these in detail, but refer to papers that give full details on the algorithms:

- **Coherent Wave Burst (CWB)**:
  - S. Klimenko et al. (2016), *Phys. Rev.* D **93**, 042004, arXiv:1511.05999.
- **MBTA**:
  - Adams, T., Buskulic, D., Germain, V., et al. (2016), *Class. Quantum Grav.* **33**, 175012, doi: 10.1088/0264-9381/33/17/175012
- **SPIIR**:
  - Luan, J., Hooper, S., Wen, L., & Chen, Y. (2012), *Phys. Rev.* D **85**, 102002, doi: 10.1103/PhysRevD.85.102002
  - Hooper, S., Chung, S. K., Luan, J., et al. (2012), *Phys. Rev.* D **86**, 024012, doi: 10.1103/PhysRevD.86.024012

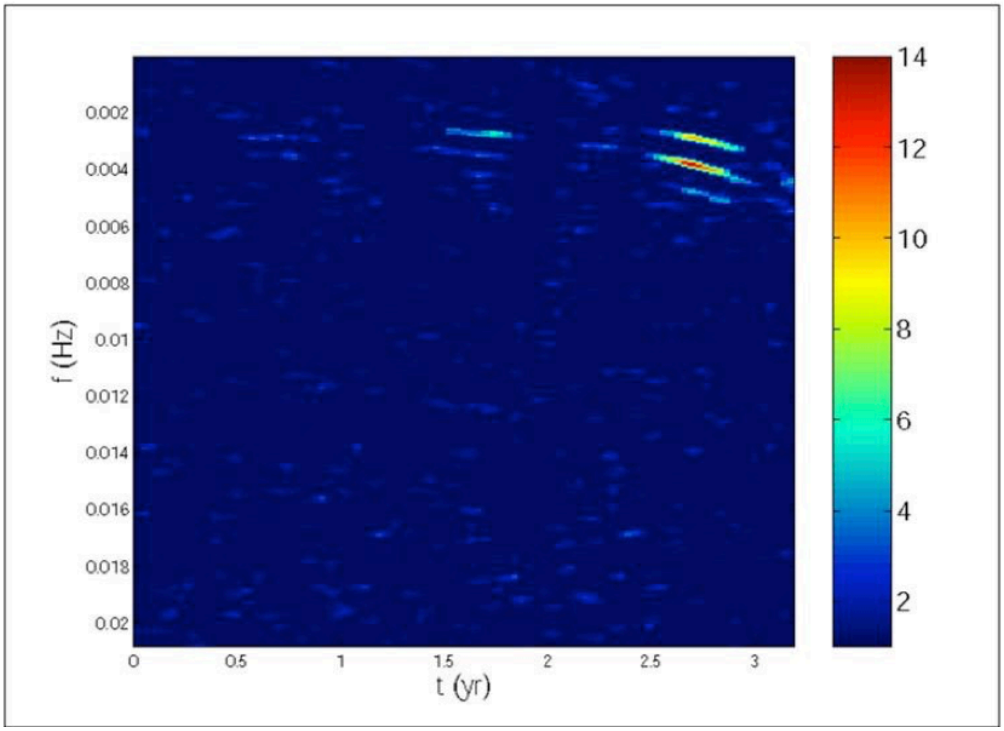


Figure 27: Example of a time-frequency spectrogram. Reproduced from Wen & Gair (2005).

- Chu, Q. (2017), PhD thesis, University of Western Australia
- Guo, X., Chu, Q., Chung, S. K., et al. 2018, *Co. Phys.* C **231**, 62, doi: 10.1016/j.cpc.2018.05.002

- **X-pipeline**:
  - Sutton, P. J., Jones, G., Chatterji, S., et al. (2010), *N J Phys.* **12**, 053034
  - Was, M., Sutton, P. J., Jones, G., & Leonor, I. (2012), *Phys. Rev.* D **86**, 022003

All of these algorithms search for clusters in **time-frequency spectrograms** of the data. The full data stream is divided into (usually overlapping) time segments, windowed and Fourier-transformed to obtain a frequency-domain representation of that chunk of data. The norm of these spectra is computed and they are then arranged next to one another in a grid. An example of a spectrogram is shown in Figure 27. Real astrophysical sources tend to produce coherent groups of bright pixels, or tracks, in these spectrograms. The patterns will be similar in different detectors in the network. The various time-frequency algorithms typically first evaluate bright pixels in the spectrograms, by thresholding on the power or some derived quantities. Then they cluster the pixels into groups, apply consistency criteria for the location of groups in two or more detectors in the network, and hence identify triggers of interest.

Time-frequency methods have also been applied to analysis of simulated LISA data, in the context of the LISA Mock Data Challenges (e.g., Gair, J.R. and Jones, G.J. (2007), *Class. Quantum Grav.* **24**, 1145; Gair, J.R., Mandel, I. and Wen, L. (2008), *Class. Quantum Grav.* **25**, 184031; Gair, J.R. and Wen, L. (2005), *Class. Quantum Grav.* **22**, S1359; Wen, L. and Gair, J.R. (2005), Detecting extreme mass ratio inspirals with LISA using time-frequency methods, *Class. Quantum Grav.* **22**, S445.). While these algorithms were successful in simplified situations (i.e., with many fewer sources in the data than we would expect to see in practice) they are unlikely to be very effective when applied to real LISA data, due to the very large number of expected sources that will be overlapping in both time and frequency.

7.5 Semi-coherent searches

For continuous gravitational wave signals, e.g., rotating neutron stars in LIGO data, or very long-lived inspiral signals, e.g., extreme-mass-ratio inspirals in LISA data, matched filtering is possible in the sense that templates of the signals can be generated. however, it is computationally impossible, because the number of templates required to ensure a dense coverage of parameter space is extremely large. In these cases, it is possible to use **semi-coherent** search methods. These involve dividing the data stream into shorter segments, analysing each of those segments with matched filtering, and them adding up the power in the matched filter outputs along trajectories through the segments that correspond to physical inspirals. This approach is summarised in Figure 28. The semi-coherent approach is more computationally efficient, because the number of templates required to densely cover the parameter space for shorter observation times is much smaller.

A discussion of the use of a semi-coherent technique for detection of extreme-mass-ratio inspirals may be found in Gair, J.R. et al. (2004), *Class. Quantum Grav.* **21**, S1595. In that context, the coherent phase used 2 week segments of data, out of 1 year long LISA data sets. The coherent phase also employs the $\mathcal{F}$-statistic described above to automatically maximize over some of the extrinsic parameters.The impact of using the semi-coherent method rather than fully coherent matched filtering is to increase the estimated matched-filtering signal-to-noise ratio threshold for detection from $\rho = 14$ to $\rho = 30$.

In the context of the ground-based detectors, similar methods are used to search for continuous gravitational wave signals from rotating pulsars. The most recent LIGO results from the O2 science run are described in this paper

- Abbott, B.P. et al. (2019), *All-sky search for continuous gravitational waves from isolated neutron stars using Advanced LIGO O2 data*, *Phys. Rev.* D **100**, 024004.

LIGO uses two primary search methods. The **time-domain F-statistic** uses the same technique as the EMRI search described above. In fact, the latter was based on the former. Further details can be found in

- Aasi, J. et al. (2014), *Class. Quantum Grav.* **31**, 165014
- Jaranowski, P., Królak, A. and Schutz, B.F. (1998), *Phys. Rev.* D **58**, 063001
- Astone, P., Borkowski, K.M., Jaranowski, P., Pietka M. and Królak, A. (2010), *Phys. Rev.* D **82**, 022005
- Pisarski, A. and Jaranowski, P. (2015), *Class. Quantum Grav.* **32**, 145014

LIGO also employs a second method, called **the Hough transform**. The first stage of this algorithm is the same as the stack-slide method, i.e., coherent matched filtering on shorter segments of data. The second stage is slightly different, using the Hough transform, which is a technique for edge-detection in images, to identify tracks through the coherent template overlaps that might correspond to true signals. Further details can be found in

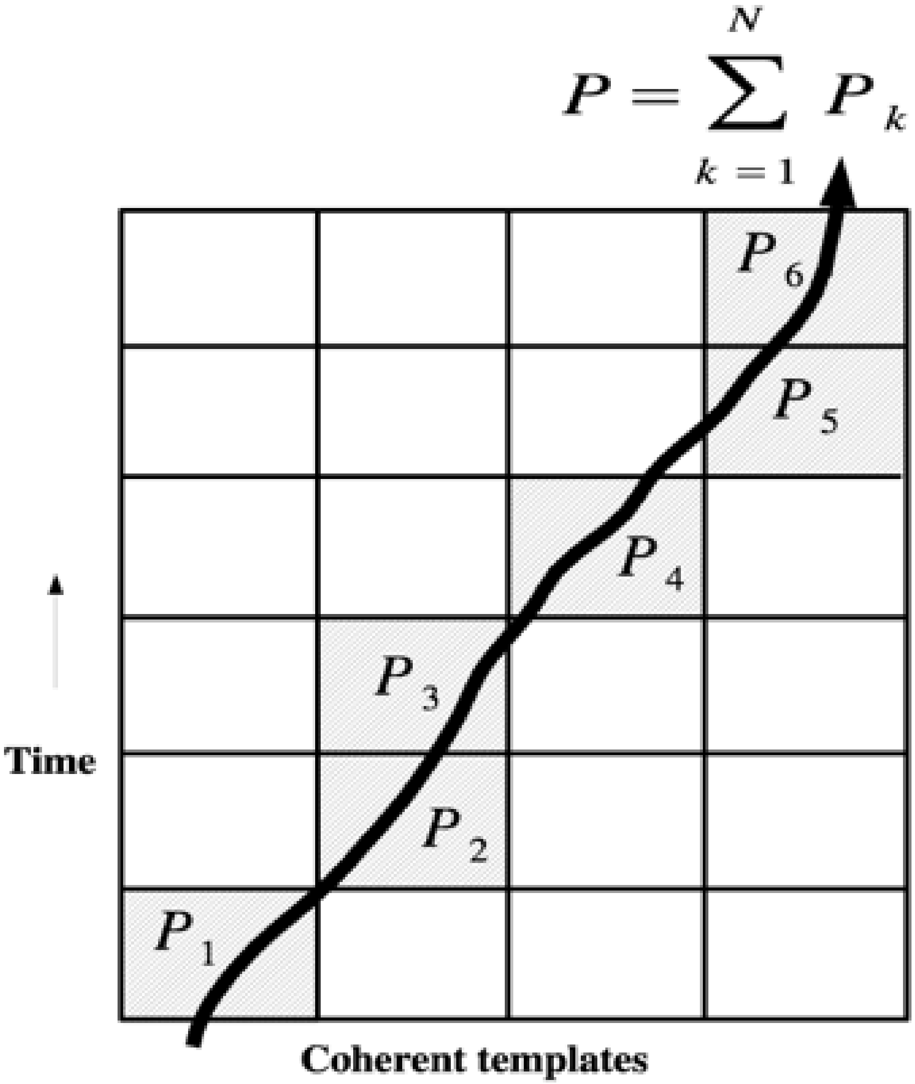


Figure 28: Illustration of the semi-coherent search method. The data is divided into shorter segments, which are searched coherently using waveform templates. The power in the templates is then summed incoherently along trajectories through the templates that correspond to EMRI inspiral trajectories. Reproduced from Gair et al. (2005).

- Astone, P., Colla, A., D?Antonio, S., Frasca, S. and Palomba, C. (2014), *Phys. Rev.* D **90**, 042002
- Antonucci, F., Astone, P., D?Antonio, S., Frasca, S. and Palomba, C. (2008), *Class. Quantum Grav.* **25**, 184015
- Krishnan, B., Sintes, A.M., Papa, M.A., Schutz, B.F., Frasca, S. and Palomba, C. (2004), *Phys. Rev.* D **70**, 082001

7.6 Searches for stochastic backgrounds

Stochastic backgrounds require different search techniques again. It is difficult to identify a background in a single detector, as it is essentially a noise source which is therefore challenging to distinguish from instrumental noise. Instead, background searches make use of multiple detectors and cross-correlate them to identify the common component of the noise. A typical detection statistic takes the form

$$\begin{aligned} Y_Q &= \int_0^T \mathrm{d}t_1 \int_0^T \mathrm{d}t_2 \; s_1(t_1) Q(t_1 - t_2) s_2(t_2) \\ &= \int_{-\infty}^{\infty} \mathrm{d}f \int_{-\infty}^{\infty} \mathrm{d}f' \, \delta_T(f - f') \tilde{s}_1^*(f) \tilde{Q}(f') \tilde{s}_2(f'). \end{aligned} \tag{105}$$

In the above, $Q(t)$ is a filter, which is analogous to the filter introduced in the single source detection case discussed earlier. The function $\delta_T(f)$ is a finite time approximation to the Dirac delta function

$$\delta_T(f) = \int_{-T/2}^{T/2} \mathrm{e}^{-2\pi i f t} \mathrm{d}t = \frac{\sin(\pi f T)}{\pi f}.$$

A generic gravitational wave background can be decomposed into a superposition of plane waves and a sum over polarisation states

$$h_{ij}(t, \vec{x}) = \int_{-\infty}^{\infty} \mathrm{d}f \int_{S^2} \mathrm{d}^{\Omega}_{\hat{k}} \mathrm{e}^{2\pi i f(t - \hat{k}\cdot\vec{x})} \mathcal{H}_A(f, \hat{k}) \mathbf{e}^A_{ij}(\hat{k}).$$

Here A labels the polarisation state, which for gravitational waves in general relativity is either plus or cross, $A = \{+, \times\}$, but in general metric theories could also include scalar and vector modes. The quantities $\mathbf{e}^A_{ij}(\hat{k})$ are the polarisation basis tensors for the individual polarisation modes

$$\mathbf{e}^+_{ij}(\hat{k}) = \hat{l}_i \hat{l}_j - \hat{m}_i \hat{m}_j, \qquad \mathbf{e}^\times_{ij}(\hat{k}) = \hat{l}_i \hat{m}_j + \hat{m}_i \hat{l}_j$$

where

$$\begin{aligned} \hat{k} &= \sin\theta\cos\phi\,\hat{x} + \sin\theta\sin\phi\,\hat{y} + \cos\theta\,\hat{z} \\ \hat{l} &= \cos\theta\cos\phi\,\hat{x} + \cos\theta\sin\phi\,\hat{y} - \sin\theta\,\hat{z} \\ \hat{m} &= -\sin\phi\,\hat{x} + \cos\phi\,\hat{y} \end{aligned} \tag{106}$$

are the standard spherical-polar coordinate basis vectors on the sky at colatitude θ and longitude ϕ. The quantities $\mathcal{H}^A(f, \hat{k})$ are the amplitudes of the various modes. For an unpolarised, stationary and statistically isotropic gravitational wave background, the expectation value of pairs of these amplitudes is given by

$$\left\langle \mathcal{H}^A(f, \hat{k}) \mathcal{H}^{A'*}(f', \hat{k}') \right\rangle = H(f) \delta(f - f') \delta^2(\hat{k}, \hat{k}') \delta_{AA'}, \tag{107}$$

where $H(f)$ is a real-valued function that depends on the energy density in the gravitational wave background and can be related to $\Omega_{\rm GW}(f)$, as introduced in the previous chapter, by

$$H(f) = \frac{3H_0^2}{32\pi^3}\frac{\Omega_{\rm GW}(f)}{|f|^3}.$$

The response of a particular gravitational wave detector, labelled by I, to a gravitational wave field can be written in the form

$$\begin{aligned} s_I(t) &= \int_{-\infty}^{\infty} \mathrm{d}\tau \int_{R^3} \mathrm{d}^3\vec{y}\, h_{ij}(t-\tau, \vec{x}-\vec{y}) R_I^{ij}(\tau, \vec{y}) \\ &= (2\pi)^3 \int_{-\infty}^{\infty} \mathrm{d}f \int_{R^3} \mathrm{d}^3\vec{k}\, \tilde{h}_{ij}(f, \vec{k}) \tilde{R}_I^{ij}(f, \vec{k}) \mathrm{e}^{i(2\pi f t - \vec{k}\cdot\vec{x}_I}) \end{aligned} \tag{108}$$

where $R^{ij}(t, \vec{x})$ is the impulse response of the detector, and the integral is over the spatial extent of the detector. Combining Eq. (108) with Eq. (107) we obtain

$$\langle Y_Q \rangle = \frac{T}{2}\int_{-\infty}^{\infty} \gamma_{12}(|f|)\tilde{Q}(f)H(f)\mathrm{d}f$$

where $\gamma(|f|)$ is the **overlap reduction function**, which depends on the relative separation and orientation of the two detectors and is defined by

$$\gamma_{12}(|f|) = \int_{S^2} \mathrm{d}\Omega_{\hat{k}}\, \tilde{R}_1^A(f, \hat{k})\tilde{R}_2^{A*}(f, \hat{k})\mathrm{e}^{-2\pi i f \hat{k}\cdot(\vec{x}_1 - \vec{x}_2)}$$

where

$$\tilde{R}_I^A(f, \hat{k}) = (2\pi)^e \mathbf{e}_{ij}^A(\hat{k})\tilde{R}_I^{ij}(f, 2\pi f\hat{k}).$$

The overlap reduction function for various combinations of ground-based interferometers and resonant bar detectors is shown in Figure 29. Stochastic backgrounds generated by large numbers of supermassive black hole binary inspirals are also the primary source for pulsar timing arrays. In that case, the “detector” is the measured redshift of a pulsar. The overlap reduction function for the detection of an isotropic stochastic background by cross-correlation of the measured redshifts of two different pulsars must be a function of only the angular separation between the pulsars on the sky. The resulting overlap reduction function curve is called the Hellings and Downs curve and is shown in Figure 30. Overlap reduction functions for non-isotropic backgrounds, for example anisotropic or correlated backgrounds, of backgrounds with non-GR polarisations, look different, providing a diagnostic for these physical properties of any observed stochastic background.

As in the case of the optimal filter, it is possible to maximise the signal-to-noise ratio of the filtered output. This takes a similar form to the optimal filter result

$$\tilde{Q}(f) \propto \frac{\gamma(|f|)\Omega_{\rm GW}(|f|)}{|f|^3 S_1(|f|)S_2(|f|)}$$

where $S_1(|f|)$ and $S_2(|f|)$ are the power spectral densities of the noise in the two detectors.

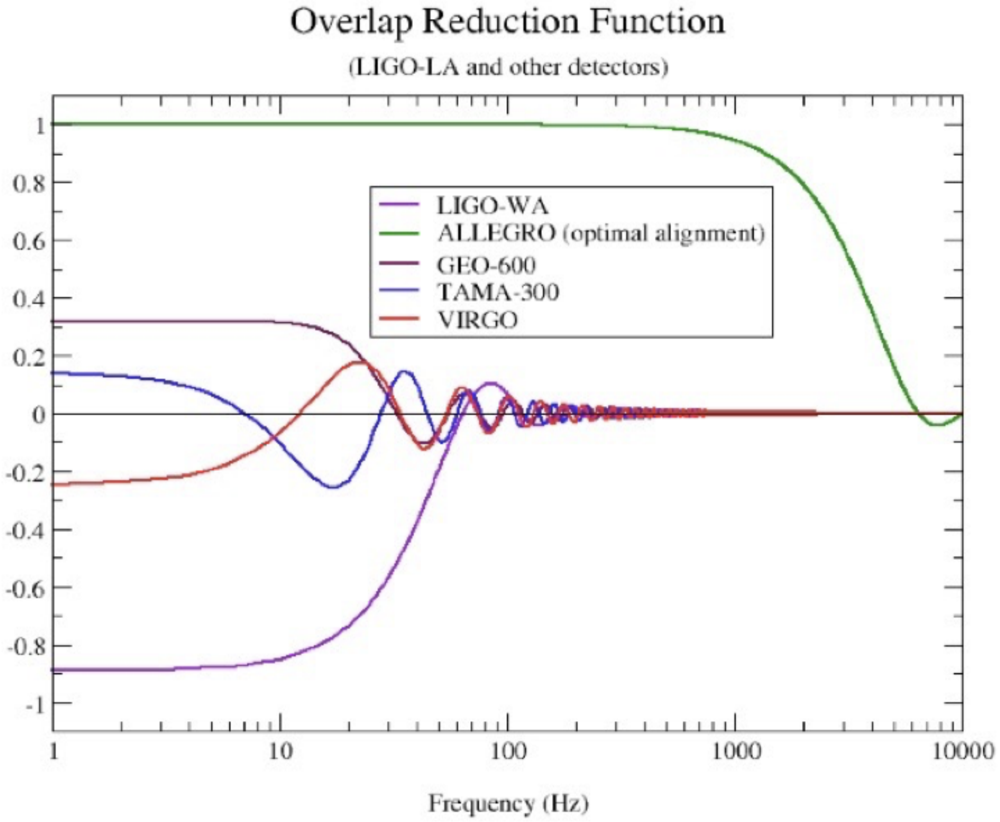


Figure 29: Overlap reduction function of the LIGO Livingston detector with LIGO Hanford (lower purple curve), Virgo (red curve), GEO (upper purple curve), TAMA (now obsolete) (blue curve) and the resonant bar detector Allegro (green curve), which was also sited in Louisiana. This was the network of detectors operating at the time of initial LIGO's science runs.

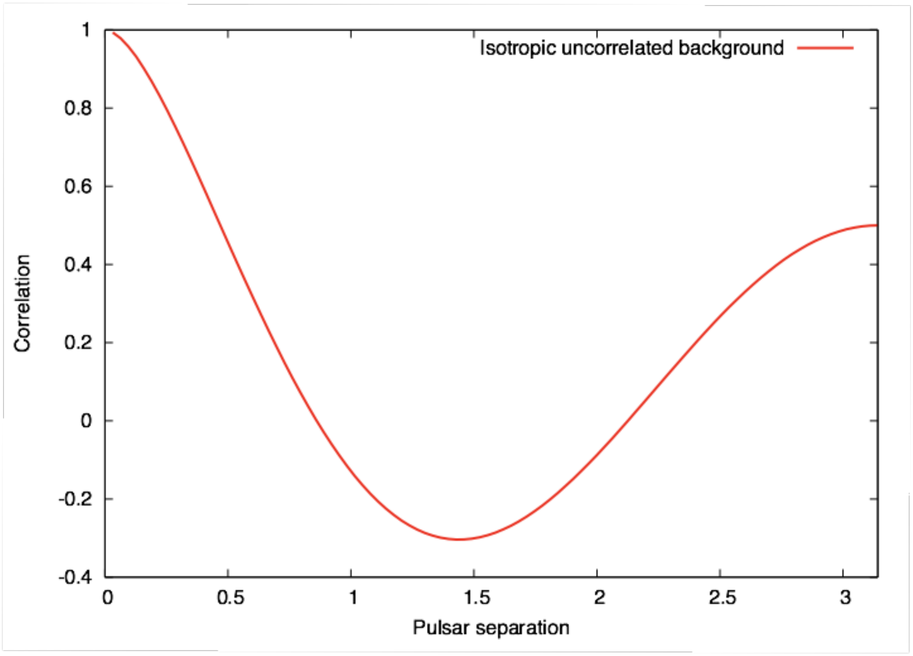


Figure 30: Overlap reduction function for the cross-correlation of the redshifts of two pulsars observed in a pulsar timing array, as a function of the angular separation of the two pulsars on the sky. This is known as the Hellings and Downs curve and the observation of a cross-correlation pattern that matches with this expectation is critical for the pulsar timing detection of gravitational waves.

8 Examples of Bayesian statistics in gravitational wave astronomy

In this section we will provide some examples of the application of Bayesian statistics in gravitational wave astronomy. In most cases we will briefly outline what is done, and provide references where further information can be obtained.

8.1 LIGO Parameter Estimation

Parameter estimation results for sources detected by the LIGO interferometers are obtained and summarised as posterior distributions using the Bayesian techniques described earlier in this course. Typically, LIGO parameter estimation results are quoted as posterior medians and symmetric credible intervals. Figure 31 gives an example of this, showing the summary of parameter estimation results for all of the events observed by LIIGO and Virgo during the O1 and O2 observing runs (Abbott et al. (2019), *Phys. Rev.* X **9** 031040).

LIGO/Virgo parameter estimation results in O1 and O2 were computed using the *LALInference* software suite, which includes two separate parameter estimation codes. *LALInferenceMCMC* is a Markov Chain Monte Carlo code, which generates posterior distributions using the Metropolis-Hastings algorithm and proposal distributions that are tuned to features expected in the likelihood for gravitational wave observations of compact binary inspirals. Further details can be found in

- Röver, C., Meyer, R., and Christensen, N., *Bayesian Inference on Compact Binary Inspiral Gravitational Radiation Signals in Interferometric Data*, *Class. Quantum Grav.* **23**, 4895 (2006).
- van der Sluys, M., Raymond, V., Mandel, I., Röver, C., Christensen, N., Kalogera, V., Meyer, R., and Vecchio, A., *Parameter Estimation of Spinning Binary Inspirals Using Markov-Chain Monte Carlo*, *Class. Quantum Grav.* **25**, 184011 (2008).

LALInferenceNest is a nested sampling algorithm, which obtains candidate values for updates to the live point set by carrying out short MCMC chains originating at the current lowest likelihood point in the live point set. Further details can be found in

- Veitch, J., and Vecchio, A., *Phys. Rev.* D **81**, 062003 (2010).

A summary of the *LALInference* package can be found in

- Veitch, J., et al., *Parameter Estimation for Compact Binaries with Ground-Based Gravitational-Wave Observations Using the LALInference Software Library*, *Phys. Rev.* D **91**, 042003 (2015).

and the version used in the analysis of the O2 events can be downloaded from

- https://git.ligo.org/lscsoft/lalsuite/tree/lalinference_o2 .

From O3 onwards, an additional parameter estimation code, *Bilby*, has been developed and used to obtain posterior distributions for LIGO/Virgo detections. This code uses generic freely available Bayesian sampling codes to draw samples from the posterior distribution, such as DYNESTY and PTMCMC. The rest of the code consists of wrappers and functions to compute the correct likelihood to feed to the sampling codes. The description of the software can be found in

Event	$m_1/M_\odot$	$m_2/M_\odot$	$\mathcal{M}/M_\odot$	$\chi_{\rm eff}$	$M_f/M_\odot$	a_f	$E_{\rm rad}/(M_\odot c^2)$	$\ell_{\rm peak}/({\rm erg\,s^{-1}})$	$d_L/{\rm Mpc}$	z	$\Delta\Omega/{\rm deg}^2$
GW150914	$35.6^{+4.7}_{-3.1}$	$30.6^{+3.0}_{-4.4}$	$28.6^{+1.7}_{-1.5}$	$-0.01^{+0.12}_{-0.13}$	$63.1^{+3.4}_{-3.0}$	$0.69^{+0.05}_{-0.04}$	$3.1^{+0.4}_{-0.4}$	$3.6^{+0.4}_{-0.4} \times 10^{56}$	440^{+150}_{-170}	$0.09^{+0.03}_{-0.03}$	182
GW151012	$23.2^{+14.9}_{-5.5}$	$13.6^{+4.1}_{-4.8}$	$15.2^{+2.1}_{-1.2}$	$0.05^{+0.31}_{-0.20}$	$35.6^{+10.8}_{-3.8}$	$0.67^{+0.13}_{-0.11}$	$1.6^{+0.6}_{-0.5}$	$3.2^{+0.8}_{-1.7} \times 10^{56}$	1080^{+550}_{-490}	$0.21^{+0.09}_{-0.09}$	1523
GW151226	$13.7^{+8.8}_{-3.2}$	$7.7^{+2.2}_{-2.5}$	$8.9^{+0.3}_{-0.3}$	$0.18^{+0.20}_{-0.12}$	$20.5^{+6.4}_{-1.5}$	$0.74^{+0.07}_{-0.05}$	$1.0^{+0.1}_{-0.2}$	$3.4^{+0.7}_{-1.7} \times 10^{56}$	450^{+180}_{-190}	$0.09^{+0.04}_{-0.04}$	1033
GW170104	$30.8^{+7.3}_{-5.6}$	$20.0^{+4.9}_{-4.6}$	$21.4^{+2.2}_{-1.8}$	$-0.04^{+0.17}_{-0.21}$	$48.9^{+5.1}_{-4.0}$	$0.66^{+0.08}_{-0.11}$	$2.2^{+0.5}_{-0.5}$	$3.3^{+0.6}_{-1.0} \times 10^{56}$	990^{+440}_{-430}	$0.20^{+0.08}_{-0.08}$	921
GW170608	$11.0^{+5.5}_{-1.7}$	$7.6^{+1.4}_{-2.2}$	$7.9^{+0.2}_{-0.2}$	$0.03^{+0.19}_{-0.07}$	$17.8^{+3.4}_{-0.7}$	$0.69^{+0.04}_{-0.04}$	$0.9^{+0.0}_{-0.1}$	$3.5^{+0.4}_{-1.3} \times 10^{56}$	320^{+120}_{-110}	$0.07^{+0.02}_{-0.02}$	392
GW170729	$50.2^{+16.2}_{-10.2}$	$34.0^{+9.1}_{-10.1}$	$35.4^{+6.5}_{-4.8}$	$0.37^{+0.21}_{-0.25}$	$79.5^{+14.7}_{-10.2}$	$0.81^{+0.07}_{-0.13}$	$4.8^{+1.7}_{-1.7}$	$4.2^{+0.9}_{-1.5} \times 10^{56}$	2840^{+1400}_{-1360}	$0.49^{+0.19}_{-0.21}$	1041
GW170809	$35.0^{+8.3}_{-5.9}$	$23.8^{+5.1}_{-5.2}$	$24.9^{+2.1}_{-1.7}$	$0.08^{+0.17}_{-0.17}$	$56.3^{+5.2}_{-3.8}$	$0.70^{+0.08}_{-0.09}$	$2.7^{+0.6}_{-0.6}$	$3.5^{+0.6}_{-0.9} \times 10^{56}$	1030^{+320}_{-390}	$0.20^{+0.05}_{-0.07}$	308
GW170814	$30.6^{+5.6}_{-3.0}$	$25.2^{+2.8}_{-4.0}$	$24.1^{+1.4}_{-1.1}$	$0.07^{+0.12}_{-0.12}$	$53.2^{+3.2}_{-2.4}$	$0.72^{+0.07}_{-0.05}$	$2.7^{+0.4}_{-0.3}$	$3.7^{+0.4}_{-0.5} \times 10^{56}$	600^{+150}_{-220}	$0.12^{+0.03}_{-0.04}$	87
GW170817	$1.46^{+0.12}_{-0.10}$	$1.27^{+0.09}_{-0.09}$	$1.186^{+0.001}_{-0.001}$	$0.00^{+0.02}_{-0.01}$	≤ 2.8	≤ 0.89	≥ 0.04	$\geq 0.1 \times 10^{56}$	40^{+7}_{-15}	$0.01^{+0.00}_{-0.00}$	16
GW170818	$35.4^{+7.5}_{-4.7}$	$26.7^{+4.3}_{-5.2}$	$26.5^{+2.1}_{-1.7}$	$-0.09^{+0.18}_{-0.21}$	$59.4^{+4.9}_{-3.8}$	$0.67^{+0.07}_{-0.08}$	$2.7^{+0.5}_{-0.5}$	$3.4^{+0.5}_{-0.7} \times 10^{56}$	1060^{+420}_{-380}	$0.21^{+0.07}_{-0.07}$	39
GW170823	$39.5^{+11.2}_{-6.7}$	$29.0^{+6.7}_{-7.8}$	$29.2^{+4.6}_{-3.6}$	$0.09^{+0.22}_{-0.26}$	$65.4^{+10.1}_{-7.4}$	$0.72^{+0.09}_{-0.12}$	$3.3^{+1.0}_{-0.9}$	$3.6^{+0.7}_{-1.1} \times 10^{56}$	1940^{+970}_{-900}	$0.35^{+0.15}_{-0.15}$	1666

Figure 31: Parameter estimation results summary from the first Gravitational Wave Transient Catalogue published by the LIGO/Virgo collaboration (*Phys. Rev.* X **9** 031040 (2019)). Results are presented as the median and 90% symmetric credible interval of the Bayesian posterior distribution.

- Ashton, G., et al. (2019), *Astrophys. J. Supp.* **241**, 27

and the software can be downloaded from

- https://git.ligo.org/lscsoft/bilby

As well as providing tables summarising the median and symmetric credible intervals for the observed sources, LIGO papers typically include plots of the full Bayesian posterior distributions. These take various forms. Two-dimensional joint posterior distributions are often given for pairs of parameters that are correlated, such as the chirp mass and mass ratio or the final mass and spin of the remnant black hole produced by the merger or the sky location of the merger event. Examples of two-dimensional posterior distributions are shown in Figure 32 and Figure 33. One dimensional posteriors are often plotted as "violin plots" to allow comparison between the results for multiple events. The violin plot plots the parameter value on the y-axis and the posterior density on the x-axis, which is opposite to the usual convention. Additionally, the posterior is reflected in the y-axis so that it is symmetric about that axis for each event. The width of the resulting violin plot is proportional to the posterior probability for the corresponding value of the parameter. An example is shown in Figure 34. Posteriors in the spins of the black holes, which is fundamentally a three-dimensional quantity, are typically represented by semi-circular density plots such as those shown in Figure 35. The full 3D posterior is marginalised over the (poorly constrained) azimuthal direction of the spin, and the resulting 2D posterior is represented on a semi-circle with the spin-magnitude as the radial direction and the angle between the spin vector and the orbital angular momentum as the angular direction. The density of the colour in these plots is proportional to the posterior density for the corresponding spin vector.

LALInference is also used to obtain posterior deviations on parameters characterising deviations from general relativity, to facilitate tests of GR. More details can be found, along with results from analysis of the O1 and O2 events, in Abbott, B.P., et al., *Phys. Rev.* D **100**, 104036.

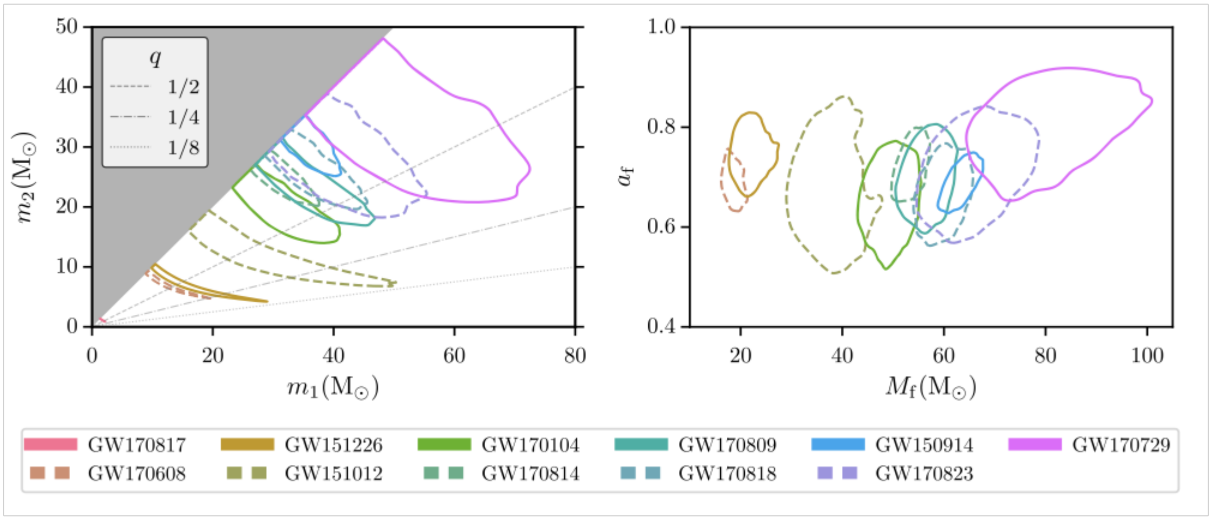


Figure 32: Joint two dimensional posterior on mass and mass ratio (left) and on final mass and spin (right) for all of the events observed by LIGO/Virgo during the O1 and O2 observing runs. Reproduced from Abbott et al. (2019), *Phys. Rev.* X **9** 031040.

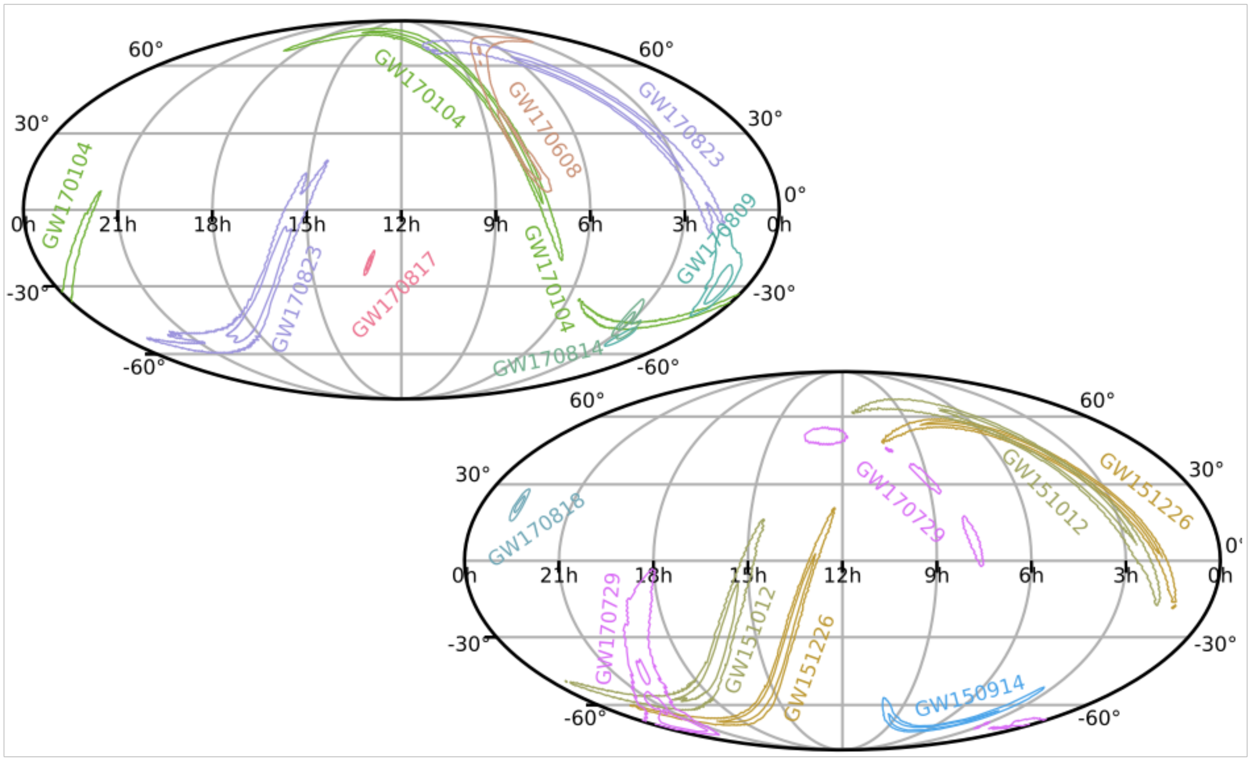


Figure 33: Sky location posterior distribution for all events observed by LIGO/Virgo during the O1 and O2 observing runs. Reproduced from Abbott et al. (2019), *Phys. Rev.* X **9** 031040.

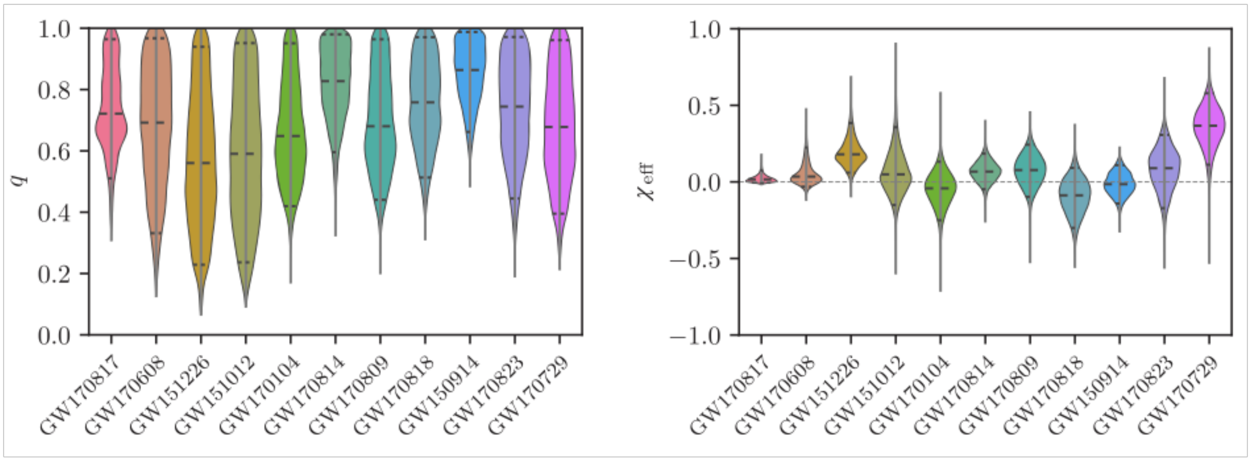


Figure 34: One-dimensional marginalised posteriors on the mass ratio (left) and effective spin (right) for all the events observed by LIGO/Virgo during the O1 and O2 observing runs. The one-dimensional posteriors are represented as "violin plots" as described in the text. Reproduced from Abbott et al. (2019), *Phys. Rev.* X **9** 031040.

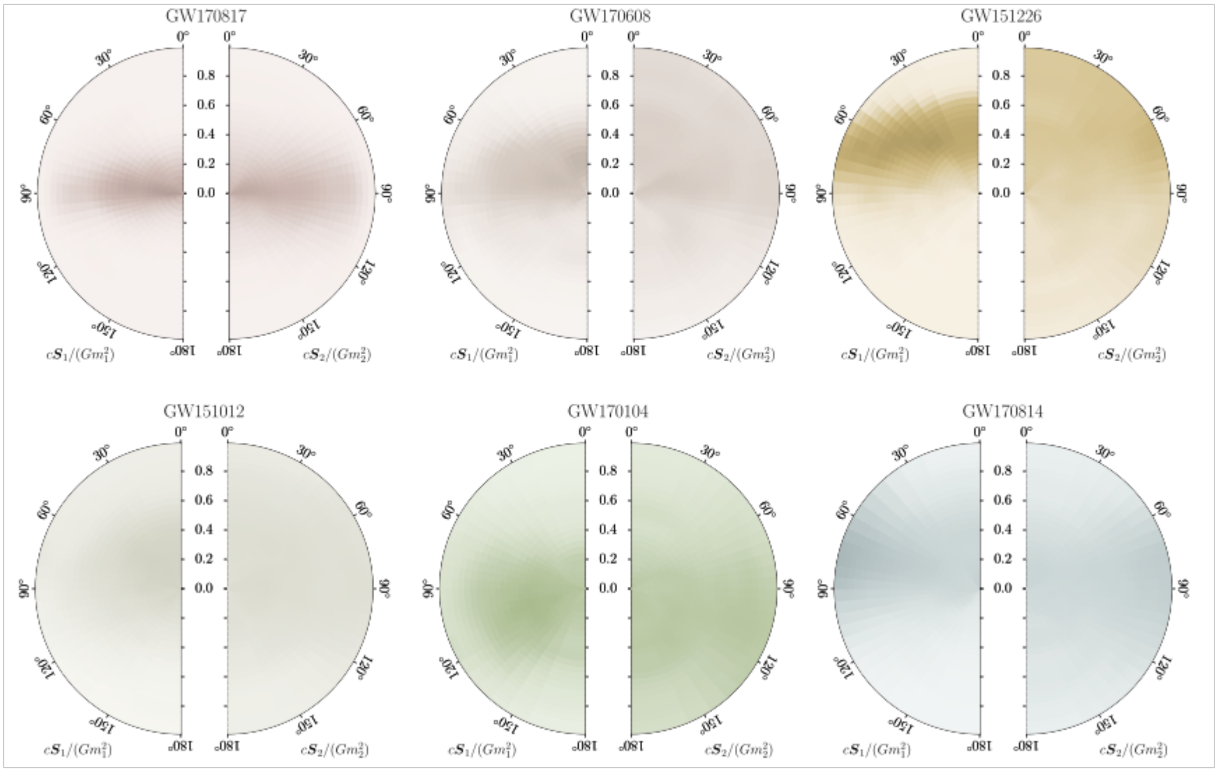


Figure 35: Posteriors on the spins of the two components in the binary for all of the events observed by LIGO/Virgo during the O1 and O2 observing runs. The distance from the origin represents the magnitude of the spin, and the angle represents the direction of the spin. The two halves of the plot are for the primary (left) and secondary (right) object in the binary. The density of colour is proportional to the posterior density for that spin value. Reproduced from Abbott et al. (2019), *Phys. Rev.* X **9** 031040.

8.2 Reduced order modelling

LIGO parameter estimation codes are computationally expensive, primarily due to the cost of evaluating models of the gravitational waveforms to compute likelihoods. To make inference more efficient, it is advantageous to have models of the signals that are quicker to evaluate. This has been achieved by building **reduced order models** and **surrogate models**. The principle of both approaches is quite similar. First, a basis for the space of waveforms is found that has lower dimensionality than the number of samples in the original waveforms. Then either a fast interpolant is constructed to map physical parameters to the weights of the basis functions (in the case of some surrogate models, the interpolant is built directly for the waveform itself) or a **reduced order quadrature** representation of the likelihood is constructed. In the latter approach, a projection of the target waveform onto the reduced basis is obtained not by using overlaps to find the best projection, but instead by requiring the target waveform to exactly match a linear combination of basis waveforms at a number of points, called **quadrature interpolation points**, equal to the number of functions in the basis. This allows the likelihood quadrature to be reduced to a sum over the target waveform evaluated at the quadrature points weighted by data-dependent constants that can be computed prior to running inference from overlaps of the basis functions with the data

$$
\begin{aligned}
\left(h(\vec{\lambda})|d\right) &= 4\Re \int_0^\infty \frac{\tilde{h}(\vec{\lambda})\tilde{d}^*(f)}{S_h(f)}\,\mathrm{d}f \\
&\approx 4\Re \left[\sum_{k=0}^{N/2} d^*(f_k)\vec{e}^T(f_k)\Delta f\,\mathbf{A}^{-1}\right]\vec{h}(\vec{\lambda}) \\
&= 4\Re \sum_{k=1}^{m} \omega_k h(F_k;\vec{\lambda}). \qquad (109)
\end{aligned}
$$

Reduced order quadrature approximations to likelihoods are the state of the art in LIGO parameter estimation, but they require being able to evaluate the target waveform at certain frequencies quickly and so can only really be used with frequency-domain waveform approximants. Surrogate models can be used to accelerate inference with time-domain waveform models.

8.3 Population inference

Inference on the properties of the population of sources form which the observed LIGO events are drawn also uses Bayesian methods, specifically Bayesian hierarchical modelling. We encountered one example of this in Section 4.9, which is the inference of cosmological parameters using gravitational wave observations of binary neutron star mergers with counterparts. Other examples include inference on the rate of mergers of different types of source in the Universe, and on the distributions of masses and spins of black holes and neutron stars. Full details on the range of population analyses carried out for the O1 and O2 events can be found in Abbott, B.P., et al., *Astrophys. J. Lett.* **882**, L24 (2019) and references therein, but we summarise some of the key analyses here.

8.3.1 Rate estimation

Accurate estimation of the rate of events in the Universe is complicated by confusion with detector noise, i.e., identifying which events are real gravitational wave events and which are instrumental artefacts, and by the need to make assumptions about the distribution of parameters of sources in the population. The first problem was tackled in Farr, W., Gair, J.R., Mandel, I., and Cutler, C., *Phys. Rev.* D **91**, 023005 (2014). If the output of the detector is represented by a sequence of values of a detection statistic, x, and any statistic value that exceeds some threshold, $x_{\rm min}$, is regarded as a detection, then the observed data is a set of detection statistic values above threshold, $\{x_i\}$. Some of these events correspond to real foreground events, while others arise due to noise fluctuations in the detector and are background. We introduce an (unobserved) parameter f_i for each event such that $f_i = 1$ is it is a foreground event and f_i=0 if it is background. The foreground and background events are assumed to be generated by independent Poisson processed with rates

$$\frac{{\rm d}N_f}{{\rm d}x} = R_f \hat{f}(x, \theta_f), \qquad \frac{{\rm d}N_b}{{\rm d}x} = R_b \hat{b}(x, \theta_b)$$

and corresponding cumulative distributions $\hat{F}(x, \theta_f)$, $\hat{B}(x, \theta_b)$. Here R_f and R_b are the foreground and background rates respectively and θ_f and θ_b represent any unknown parameters that characterise the foreground and background distributions. The combined posterior for the rates, event flags and distribution parameters is

$$p(f_i, R_f, R_b, \theta | d_{\rm to}, N) = \frac{\alpha}{p(d_{\rm to}, N) N!} \left[\prod_{i|f_i=1} R_f \hat{f}(x_i, \theta) \right] \left[\prod_{i|f_i=0} R_b \hat{b}(x_i, \theta) \right] \exp[-(R_f + R_b)] \frac{p(\theta)}{\sqrt{R_f R_b}}$$

where $p(\theta)$ is the prior on the posterior parameters and we are using a Jeffreys' prior $p(R) \propto 1/\sqrt{R}$ on the rates. The subscript on $d_{\rm to}$ indicates that we are using time-ordered data. The data could also be analysed ordered by ranking statistic. This posterior can be marginalised over the unknown flags to give posteriors on the rates, or over the rates to give posterior probabilities for $f_i = 1$ for each event.

One complication with this approach is that it relies on a model for the foreground and background distributions. These can be estimated by injections and time-slides, but, since LIGO is not equally sensitive to all types of CBC event, the former requires imposing some model of the astrophysical population from which the events are drawn. One approach to this is to assume that all events in the Universe are the same as the one that has been observed. This approach was used in Kim, Kalogera and Lorimer (*Astrophys. J.* **584**, 985 (2003)) to estimate the rate of double neutron star mergers and so is often referred to as the "KKL method". In the first LIGO detection paper, for GW150914, the combination of the rate estimation accounting for confusion (FGMC) and the KKL method was used to infer the rate of binary black hole mergers. The application of this "alphabet soup" method was complicated by the fact that the data being analysed to infer the background for GW150914 contained a second CBC trigger, LVT151012. The parameters of this event were completely different to GW150914, so the KKL method could still be applied, but generalising to the case where all events in the Universe were either like GW150914 or LVT151012. Further details can be found in Abbott, B.P., et al. *Astrophys. J. Lett.* **833**, 1 (2016) and Abbott, B.P., et al. *Astrophys. J. Supp.* **227**, 14 (2016).

One additional trigger, GW151226, was present in the LIGO O1 data, and that again had sufficiently distinct parameters that the KKL approach could be used. In O2, the events began to have much more posterior overlap and so this method could no longer be used. Now, a model of the population is assumed in event rate estimation. Recent analyses have used both a power-law mass distribution or a flat in log-mass distribution in an attempt to bound the range of possible rate, but future results are likely to shift towards a single combined analysis of the population parameters and rate.

8.3.2 Black hole mass distribution

The mass distribution of stellar-origin black holes in binaries can be inferred from LIGO/Virgo observations in a hierarchical analysis by placing a prior on the mass of individual events that depends on some unknown parameters that can be constrained from analysing the full set of events. In Abbott, B.P., et al., *Astrophys. J. Lett.* **882**, L24 (2019) three different models of the mass function were used. Models A and B assumed a power law distribution on mass and mass ratio

$$p(m_1, m_2 | m_{\min}, m_{\max}, \alpha, \beta_q) \propto \begin{cases} C(m_1) m_1^{-\alpha} q^{\beta_q} & \text{if } m_{\min} \leq m_2 \leq m_1 \leq m_{\max} \\ 0 & \text{otherwise} \end{cases} .$$

In model A, $m_{\min} = 5M_\odot$, $\beta_q = 0$ and the only free parameters are $m_{\max}$ and α. In model B, all four parameters are allowed to vary. The third model mixes a power-law component of the above form, with a Gaussian component, designed to fit any excess of events near the lower mass limit of the pair-instability supernova mass gap. The model is

$$\begin{aligned} p(m_1|\theta) &= \left[(1-\lambda_m) A(\theta) m_1^{-\alpha} \Theta(m_{\max} - m_1) + \lambda_m B(\theta) \exp\left(-\frac{(m_1 - \mu_m)^2}{2\sigma_m^2} \right) \right] S(m_1, m_{\min}, \delta m) \\ p(q = m_2/m_1 | m_1, \theta) &= C(m_1, \theta) q^{\beta_q} S(m_2, m_{\min}, \delta m). \end{aligned} \tag{110}$$

The mass distribution obtained by fitting these models to the O1 and O2 data is shown in Figure 36.

8.3.3 Black hole spin distribution

A hierarchical analysis of LIGO/Virgo events can also provide insight into the spin distribution. This can be done either parametrically or non-parametrically and both analyses were done for the O1 and O2 events in Abbott, B.P., et al., *Astrophys. J. Lett.* **882**, L24 (2019). The parametric approach models the spin magnitude using a Beta distribution

$$p(a_i | \alpha_a, \beta_a) = \frac{a_i^{\alpha_a - 1} (1 - a_i)^{\beta_a - 1}}{B(\alpha_a, \beta_a)}$$

while the non-parametric analysis models the spin-magnitude distribution as a set of heights of a binned distribution, with the bin heights free parameters to be determined by the observations. For example, a three-bin distribution (Farr, B., Holz, D., and Farr, W., *Astrophys. J.* **854**, L9 (2018))

$$p(a) = \begin{cases} A_1/3 & 0 \leq a \leq 1/3 \\ A_2/3 & 1/3 \leq a \leq 2/3 \\ 1 - (A_1 + A_2)/3 & 2/3 \leq a \leq 1 \end{cases} .$$

The posteriors obtained from applying these models to the O1 and O2 events are shown in Figure 37.

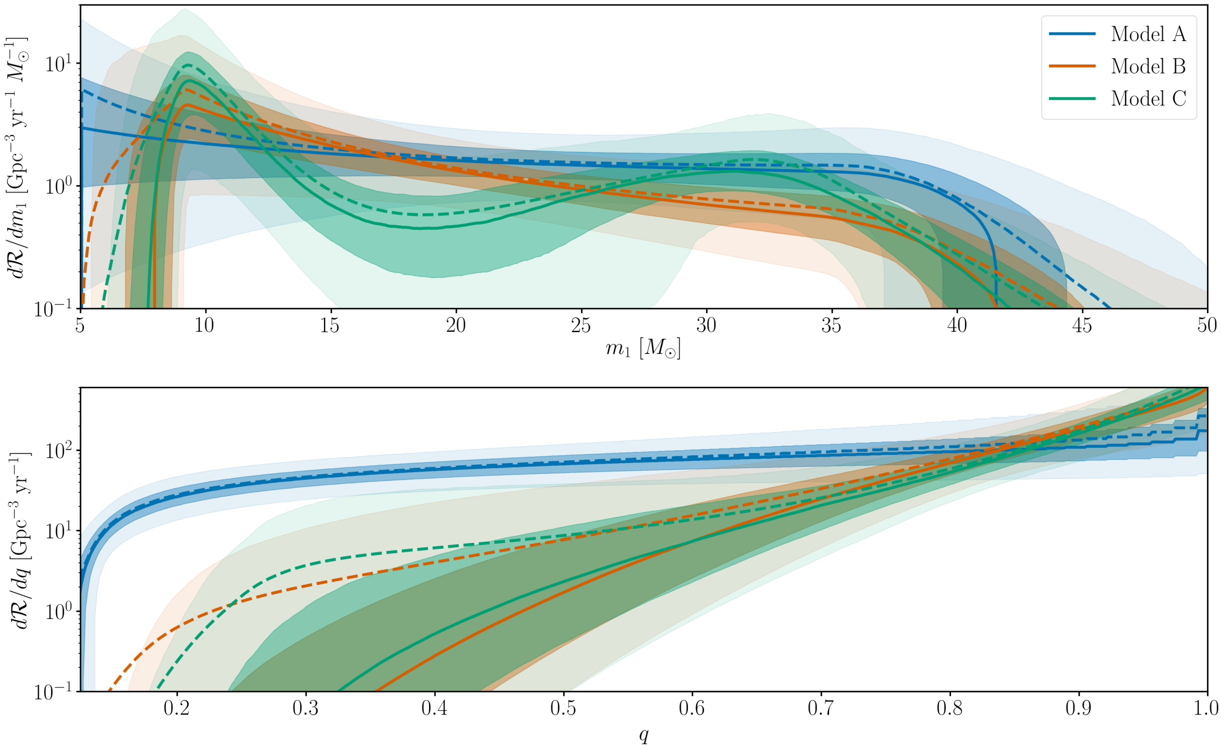


Figure 36: Black hole mass function inferred from LIGO/Virgo events observed in the O1 and O2 observing runs. Figure reproduced from Abbott, B.P., et al., *Astrophys. J. Lett.* **882**, L24 (2019).

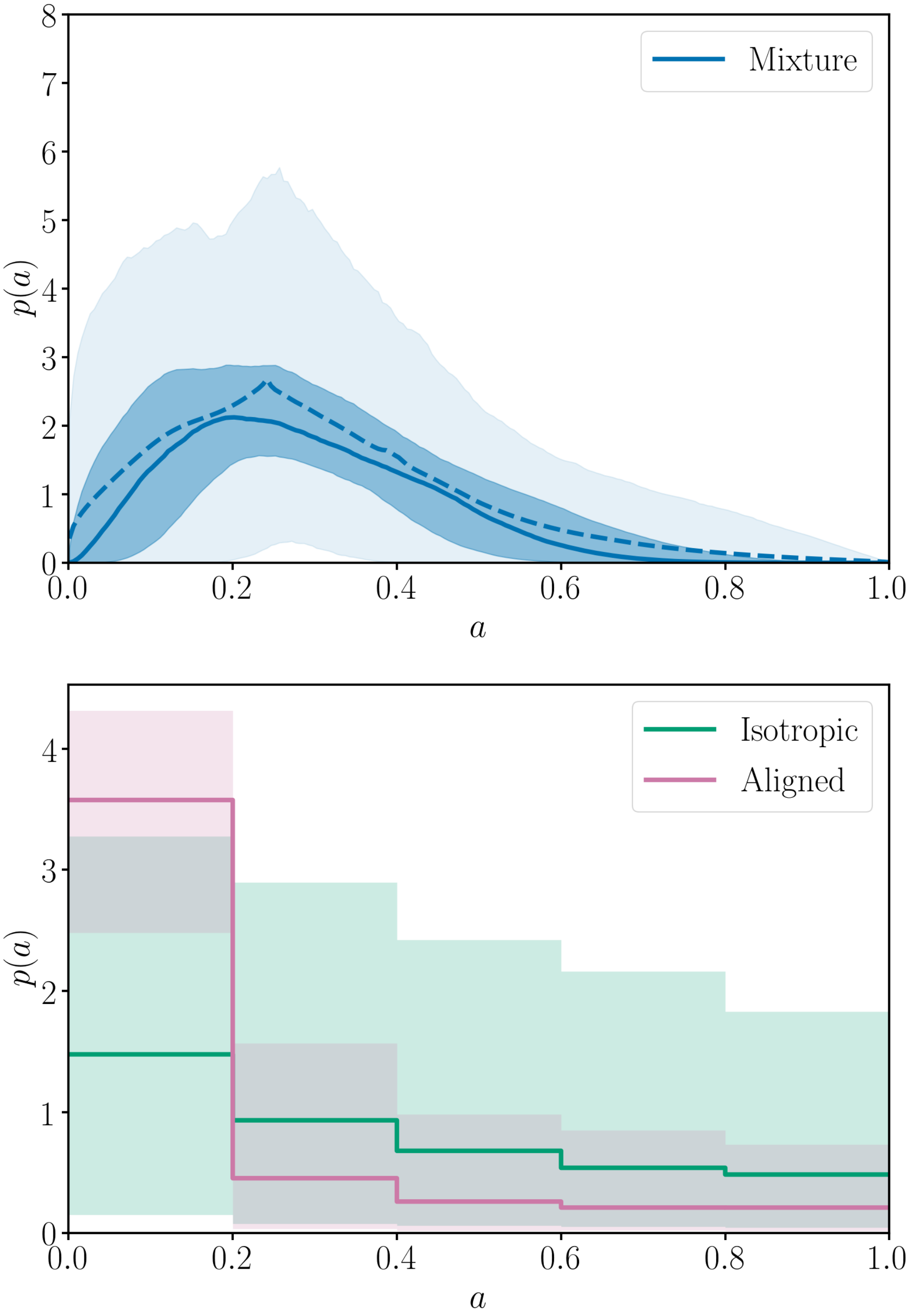


Figure 37: Black hole spin distribution inferred from LIGO/Virgo events observed in the O1 and O2 observing runs, using a parametric (top panel) or non-parametric (bottom panel) approach. Figures reproduced from Abbott, B.P., et al., *Astrophys. J. Lett.* **882**, L24 (2019).

The spin direction is also a parameter of interest astrophysically, as different formation scenarios predict either isotropically distributed spin directions, or a preference for spins to be aligned with the angular momentum of the binary. To capture this, the analysis of the O1 and O2 data used a mixture model

$$p(\cos t_1, \cos t_2 | \sigma_1, \sigma_2, \zeta) = \frac{(1-\zeta)}{4} + \frac{2\zeta}{\pi} \prod_{i \in \{1,2\}} \frac{\exp(-(1-\cos t_i)^2/2\sigma_1^2)}{\sigma_i \mathrm{erf}(\sqrt{2}/\sigma_i)}.$$

At present, LIGO measurements are not sufficiently informative about spins to strongly constrain the parameters of the model.

8.3.4 Rate evolution

The FGMC+KKL method described earlier assumes that the rate of mergers is constant, but in principle this could evolve over cosmic history (the FGMC framework can handle this, but the interpretation of R_f is different, as the average rate over the sensitive volume of the detector). An evolution of the rate can be explicitly included and constrained by introducing an extra parameter into the rate density

$$\frac{\mathrm{d}R}{\mathrm{d}\xi}(z|\theta) = R_0 p(\xi|\theta)(1+z)^\lambda.$$

The analysis of the O1 and O2 events provided weak evidence for an evolution in rate with redshift, but this was mostly due to the event GW170729, which was the most marginal detection. The rate evolution will be better constrained by the order of magnitude increase in events expected in O3 and future observing runs.

8.4 Model selection

Bayesian methods are also applied to model selection using the LIGO/Virgo observations, through the evaluation of **evidence ratios** or **Bayes factors** for pairs of alternative hypotheses for the data. Some examples of applications to gravitational wave data are

- Test for the presence of a signal in the data after the end of the merger of the two neutron stars in GW170817. Such a signal might be evidence that the merger project was a hypermassive neutron star rather than a black hole. For GW170817 the Bayes factor for the noise model over the signal model was 256.79 (Abbott, B.P. ,et al., *Phys. Rev.* X **9** 011001 (2019)), providing strong evidence that no such signal was present.

- Test of the polarisation state of gravitational waves. Possible models are that the gravitational waves have tensor polarisation, as expected in GR, or have scalar polarisation or vector polarisation. The analysis of GW170818 gave Bayes factors of 12 for tensor versus vector polarisation and 407 for tensor versus scalar, while the analysis of GW170814 gave Bayes' factors of 30 and 220 respectively (Abbott, B.P. ,et al., *Phys. Rev.* D **100** 104036 (2019)).

- Tests of the no-hair property of the remnant black hole formed in a merger, by comparing the properties of the observed ringdown radiation to that predicted by GR (Brito, Buonanno and Raymond, *Phys. Rev.* D **98**, 084038 (2018)).

- Probing alternative theories of gravity. For example, looking for evidence for dynamical gravity with the polarisation of continuous gravitational waves (Isi et al., *Phys. Rev.* D **96**, 042001 (2017)).

8.5 Source reconstruction

Although Bayesian inference relies on the existence of models, it is also possible to use these methods to recover "unmodelled" sources. One such implementation is the BayesWave algorithm. The method works by modelling the noise and signals in the data from the various detectors as a superposition of simple components. BayesWave represents the noise as a combination of a smooth PSD component, described by a cubic spline, lines represented by Lorentzians and glitches modelled by wavelets. Signals in the data are also modelled by wavelets, but with parameters that are common across the detectors, as opposed to the noise components which are independent in different detectors. Wavelets are simple functions that are compact in both time and frequency. We will encounter these again in the non-parametric regression section of this course. There are many different wavelet families, but the wavelets used in BayesWave are known as the Morley-Gabor basis.

BayesWave fits itsmodel using reversible jump MCMC. The reversible jump element is required to add or remove wavelet or line components, as the number of these required is not known a priori. Further details on the BayesWave algorithm can be found in

- Cornish, N.J., and Littenberg, T.B., *Class. Quantum Grav.* **32**, 135012 (2015).
- Littenberg, T.B., and Cornish, N.J., *Phys. Rev.* D **91**, 084034 (2015).

BayeWave is used in LIGO analyses for PSD estimation, glitch removal and for non-parametric waveform reconstruction. The good agreement between the BayesWave reconstructed waveform and the best fit model found by parameter estimation for GW150914 (see Figure 38) provided extra support to the fact that this was a true signal.

8.6 Rapid localisation

Since the start of the O1 observing run, LIGO/Virgo have been sending out triggers to facilitate follow-up of gravitational wave events by electromagnetic telescopes. To avoid delays to these alerts, it is necessary to rapidly estimate the sky location of the triggers so that astronomers know where to point their telescopes. Bayesian techniques are also used for this purpose. Full Bayesian parameter estimation is not possible in low-latency, so the rapid localisation algorithms are not truly Bayesian, but make approximations in evaluating the posterior that allow it to be computed quickly.

The Bayestar algorithm replaces the full likelihood by the autocorrelation likelihood, which is the likelihood evaluated at the maximum likelihood parameter values, as returned by the online search algorithms. This autocorrelation likelihood takes the form

$$\exp\left[-\frac{1}{2}\sum_i \rho_i^2 + \sum_i \rho_i \Re\left\{\mathrm{e}^{-i\gamma_i} z_i^*(\tau_i)\right\}\right]$$

where ρ_i denotes the signal to noise ratio in detector i, γ_i and τ_i are the phase and time of arrival of the trigger in detector i and $z_i(t)$ is the time-series of the matched filter overlap in detector i. The marginalisation of this integral over all parameters except sky location is

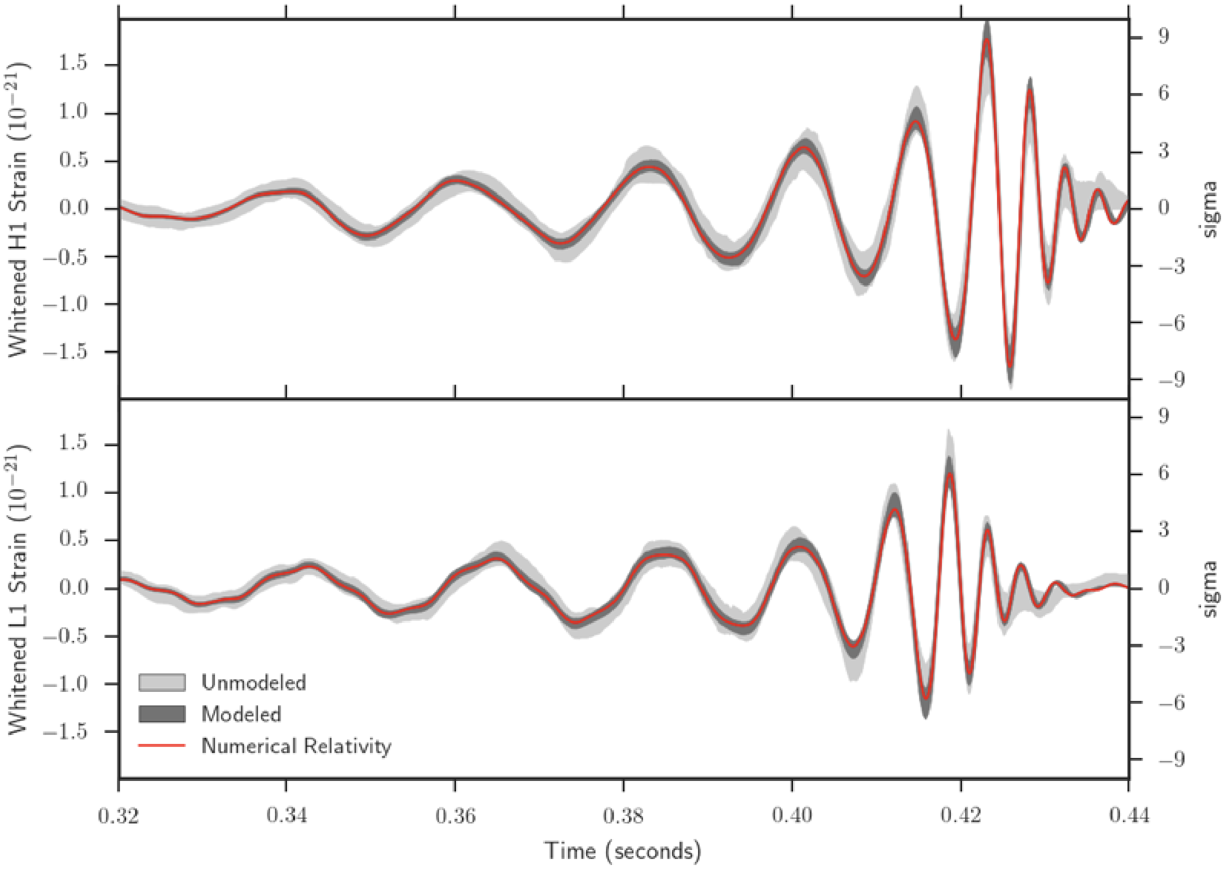


Figure 38: BayesWave reconstruction of GW150914 (labelled “unmodelled”), compared to the waveform corresponding to the maximum a posteriori parameters obtained by parameter estimation (labelled “modelled”) and a numerical relativity waveform with consistent parameters. Figure reproduced from Abbott, B.P., et al., *Phys. Rev. Lett.* **116**, 061102 (2016).

accelerated using approximations to the marginalisation integrals and by employing look-up tables. The result of running the algorithm is a sky map probability density, i.e., a weighting of pixels on the sky by their relative probability of being the true location of the observed transient.

More details on the BAYESTAR algorithm can be found in

- Singer, L., and Price, L., *Phys. Rev.* D **93**, 024013 (2016).

Another rapid localisation algorithm used in LIGO is LALINFERENCEBURST or LIB. In this case, computational savings in the model are obtained by representing an arbitrary signal as a single sine-Gaussian

$$h_+(t) = \cos(\alpha) \frac{h_{\mathrm{rss}}}{\sqrt{Q(1+\cos(2\phi_0)\mathrm{e}^{-Q^2})/4f_0\sqrt{\pi}}} \sin(2\pi f_0(t-t_0)+\phi_0)\mathrm{e}^{-(t-t_0)^2/\tau^2}.$$

While this simple model cannot accurately describe all signals, it does represent the relative amplitudes of the signal in different detectors correctly and that is enough to obtain reasonable sky-localisation accuracies.

There is also an online version of LIB, called oLIB, that uses Bayesian evidences computed by LIB to assess triggers identified in a time-frequency analysis. The evidences for the triggers being noise versus signal and being coherent in different detectors versus incoherent are used to identify potentially interesting candidate events for follow-up. oLIB was running at the time of GW150914 and, along with CWB, was the first algorithm to identify this signal in the data.

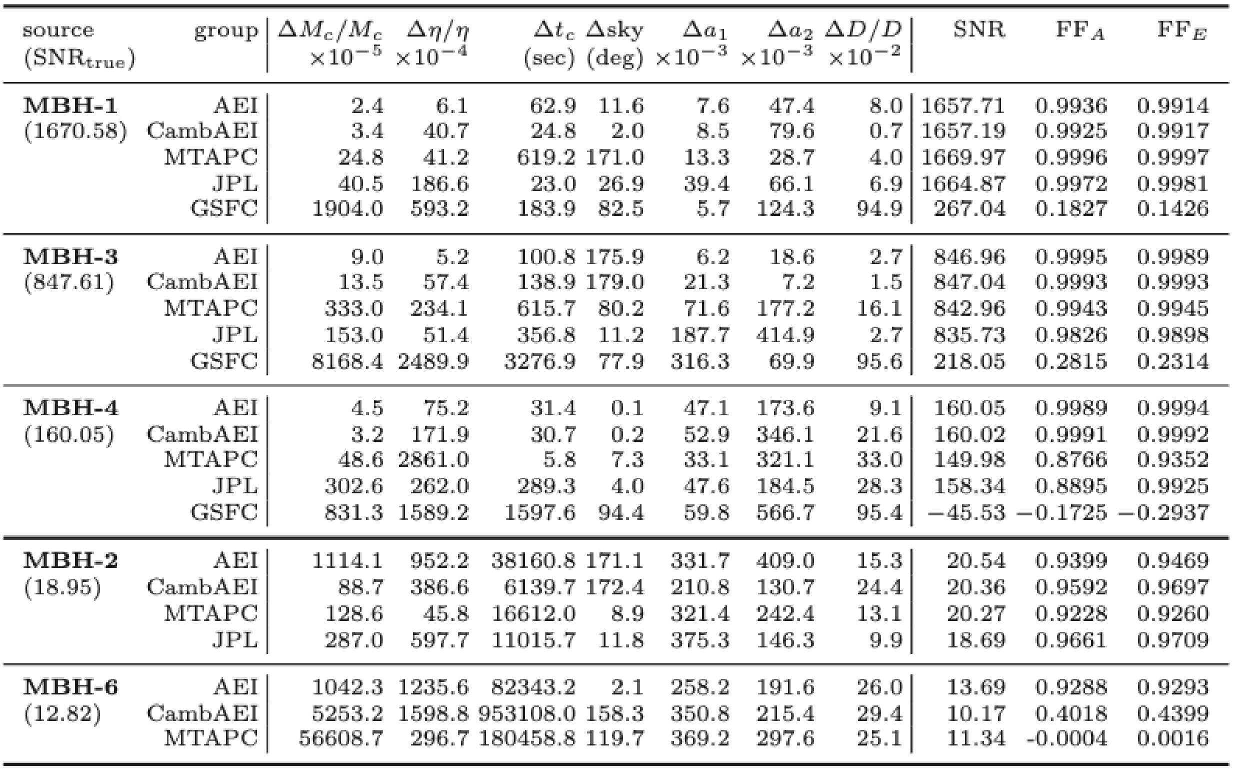
More details on the LALINFERENCEBURST algorithm and on *oLIB*, can be found in

- Essick, R., Vitale, S., Katsavounidis, E., Vedovato, G., and Klimenko, S., *Astrophys. J.* **800**, 81 (2015).
- Lynch, R., Vitale, S., Essick, R., Katsavounidis, E., and Robinet, F., *Phys. Rev.* D **95**, 104046 (2017).

8.7 LISA parameter estimation

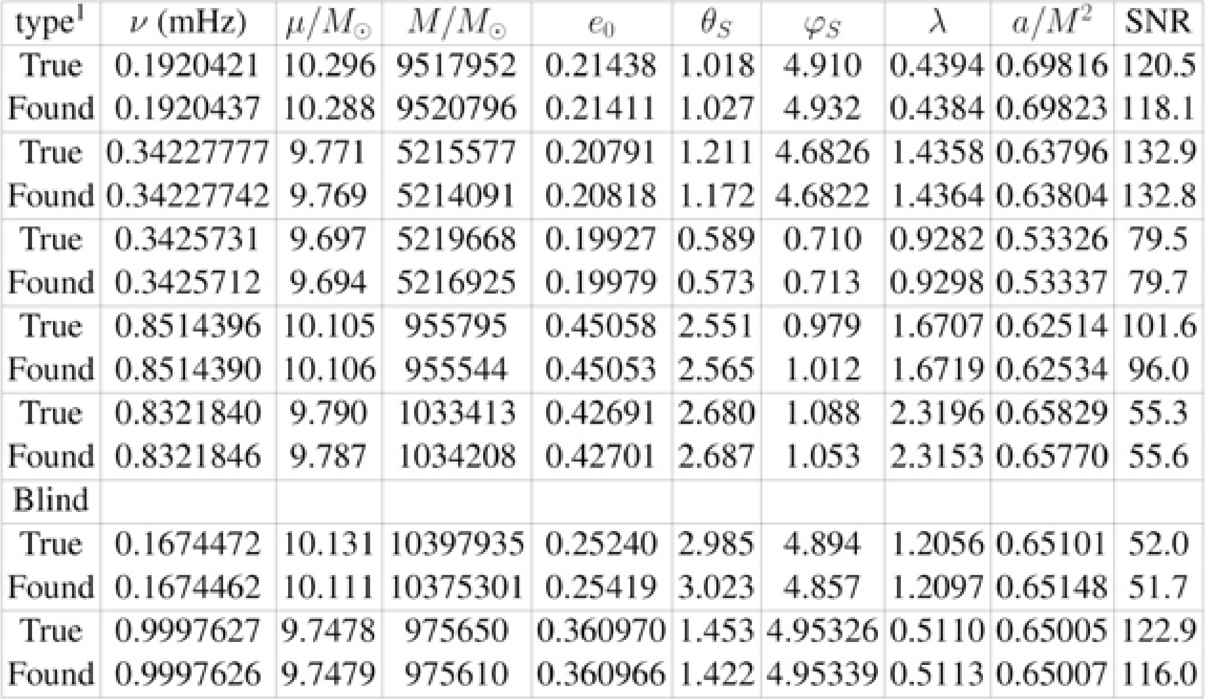
Bayesian methods have also been used in the context of data analysis development for LISA, mostly in the framework of the sequence of Mock LISA Data Challenges (MLDCs) that took place between 2006 and 2010. Bayesian techniques, with some frequentist simplifications such as the use of the $\mathcal{F}$-statistic, were used not only to characterise the identified sources, but also to search for sources in the data set. A variety of techniques were employed, including Markov Chain Monte Carlo algorithms, genetic algorithms and nested sampling. These methods were successfully able to find and characterise sources in the sample data sets, although these were somewhat simplified, containing only Gaussian instrumental noise with known PSD and a reduced number of astrophysical sources. In Figure 39 we show a table of parameter measurement precisions of supermassive black hole mergers for all submissions to the third round of the MLDC. The final two columns of the table show the fitting factor, i.e., overlap, of the submitted entry with the true source in each of the two independent LISA data channels, A and E.

The use of Bayesian techniques for searches as well as parameter estimation in the LISA context is motivated by the nature of the data. In the LIGO/Virgo context, most sources are



source ($\mathrm{SNR}_{\mathrm{true}}$)	group	$\Delta M_c/M_c$ $\times 10^{-5}$	$\Delta\eta/\eta$ $\times 10^{-4}$	Δt_c (sec)	Δsky (deg)	Δa_1 $\times 10^{-3}$	Δa_2 $\times 10^{-3}$	$\Delta D/D$ $\times 10^{-2}$	SNR	FF_A	FF_E
MBH-1	AEI	2.4	6.1	62.9	11.6	7.6	47.4	8.0	1657.71	0.9936	0.9914
(1670.58)	CambAEI	3.4	40.7	24.8	2.0	8.5	79.6	0.7	1657.19	0.9925	0.9917
	MTAPC	24.8	41.2	619.2	171.0	13.3	28.7	4.0	1669.97	0.9996	0.9997
	JPL	40.5	186.6	23.0	26.9	39.4	66.1	6.9	1664.87	0.9972	0.9981
	GSFC	1904.0	593.2	183.9	82.5	5.7	124.3	94.9	267.04	0.1827	0.1426
MBH-3	AEI	9.0	5.2	100.8	175.9	6.2	18.6	2.7	846.96	0.9995	0.9989
(847.61)	CambAEI	13.5	57.4	138.9	179.0	21.3	7.2	1.5	847.04	0.9993	0.9993
	MTAPC	333.0	234.1	615.7	80.2	71.6	177.2	16.1	842.96	0.9943	0.9945
	JPL	153.0	51.4	356.8	11.2	187.7	414.9	2.7	835.73	0.9826	0.9898
	GSFC	8168.4	2489.9	3276.9	77.9	316.3	69.9	95.6	218.05	0.2815	0.2314
MBH-4	AEI	4.5	75.2	31.4	0.1	47.1	173.6	9.1	160.05	0.9989	0.9994
(160.05)	CambAEI	3.2	171.9	30.7	0.2	52.9	346.1	21.6	160.02	0.9991	0.9992
	MTAPC	48.6	2861.0	5.8	7.3	33.1	321.1	33.0	149.98	0.8766	0.9352
	JPL	302.6	262.0	289.3	4.0	47.6	184.5	28.3	158.34	0.8895	0.9925
	GSFC	831.3	1589.2	1597.6	94.4	59.8	566.7	95.4	−45.53	−0.1725	−0.2937
MBH-2	AEI	1114.1	952.2	38160.8	171.1	331.7	409.0	15.3	20.54	0.9399	0.9469
(18.95)	CambAEI	88.7	386.6	6139.7	172.4	210.8	130.7	24.4	20.36	0.9592	0.9697
	MTAPC	128.6	45.8	16612.0	8.9	321.4	242.4	13.1	20.27	0.9228	0.9260
	JPL	287.0	597.7	11015.7	11.8	375.3	146.3	9.9	18.69	0.9661	0.9709
MBH-6	AEI	1042.3	1235.6	82343.2	2.1	258.2	191.6	26.0	13.69	0.9288	0.9293
(12.82)	CambAEI	5253.2	1598.8	953108.0	158.3	350.8	215.4	29.4	10.17	0.4018	0.4399
	MTAPC	56608.7	296.7	180458.8	119.7	369.2	297.6	25.1	11.34	-0.0004	0.0016

Figure 39: Summary of the fractional errors in the recovery of parameters of the supermassive black hole binary mergers in the third MLDC data challenge. The final two columns, labelled FF_A and FF_E, give the overlap (or "fitting factor") of the waveform corresponding to the recovered parameters with the true injected waveform. Each row represents a separate entry from one of the groups responding to the challenge. Table reproduced from Babak, S., et al., *Class. Quantum Grav.* **27**, 084009 (2010).



type^1	ν (mHz)	$\mu/M_\odot$	$M/M_\odot$	e_0	θ_S	φ_S	λ	a/M^2	SNR
True	0.1920421	10.296	9517952	0.21438	1.018	4.910	0.4394	0.69816	120.5
Found	0.1920437	10.288	9520796	0.21411	1.027	4.932	0.4384	0.69823	118.1
True	0.34227777	9.771	5215577	0.20791	1.211	4.6826	1.4358	0.63796	132.9
Found	0.34227742	9.769	5214091	0.20818	1.172	4.6822	1.4364	0.63804	132.8
True	0.3425731	9.697	5219668	0.19927	0.589	0.710	0.9282	0.53326	79.5
Found	0.3425712	9.694	5216925	0.19979	0.573	0.713	0.9298	0.53337	79.7
True	0.8514396	10.105	955795	0.45058	2.551	0.979	1.6707	0.62514	101.6
Found	0.8514390	10.106	955544	0.45053	2.565	1.012	1.6719	0.62534	96.0
True	0.8321840	9.790	1033413	0.42691	2.680	1.088	2.3196	0.65829	55.3
Found	0.8321846	9.787	1034208	0.42701	2.687	1.053	2.3153	0.65770	55.6
Blind									
True	0.1674472	10.131	10397935	0.25240	2.985	4.894	1.2056	0.65101	52.0
Found	0.1674462	10.111	10375301	0.25419	3.023	4.857	1.2097	0.65148	51.7
True	0.9997627	9.7478	975650	0.360970	1.453	4.95326	0.5110	0.65005	122.9
Found	0.9997626	9.7479	975610	0.360966	1.422	4.95339	0.5113	0.65007	116.0

Figure 40: Maximum a posteriori parameter values (labelled “Found”) recovered for all five EMRIs in the MLDC data set 1B (upper rows) and two additional random chosen sources. These are compared to the “Ture” parameters which were used ot generate the injected signals. Table reproduced from Babak, S., Gair, J.R., and Porter, E.K., *Class. Quantum Grav.***26**, 135004 (2009).

of short duration relative to the time between signals, and so it is necessary to efficiently sift through large amounts of data to find candidate sources of interest. In the LISA context, the source duration is comparable to the length of the data stream and so the entire data stream is relevant for the analysis of all sources. It is natural therefore to find and characterise sources simultaneously.

While the MLDCs demonstrated the effectiveness of the use of Bayesian methods to find and characterise most source types, several open questions remain, in particular related to the impact of non-stationary noise and instrumental artefacts such as gaps, the full extent of source confusion and the detection and characterisation of extreme-mass-ratio inspirals (EMRIs). While the EMRI sources in the MLDC data sets were successfully characterised under simplified assumptions (see Figure 41), the likelihood for an EMRI is very complicated, with many secondary maxima in parameter space. The successful algorithms relied on knowledge of the structure of the likelihood surface, which was specific to the simplified model of the EMRI employed in the MLDC, and the fact that all identified secondaries were generated by the same EMRI signal. While the structure of the likelihood surface can probably be learned for more accurate waveform models, the correct grouping of secondary modes will be much more challenging for real LISA data which could contain many hundreds of EMRIs.

Nested sampling has also been used in the context of LISA data analysis. In fact, the first application of the MultiNest nested sampling algorithm in a gravitational wave context was to the characterisation of supermassive black hole mergers in LISA data (Feroz, F., Gair,

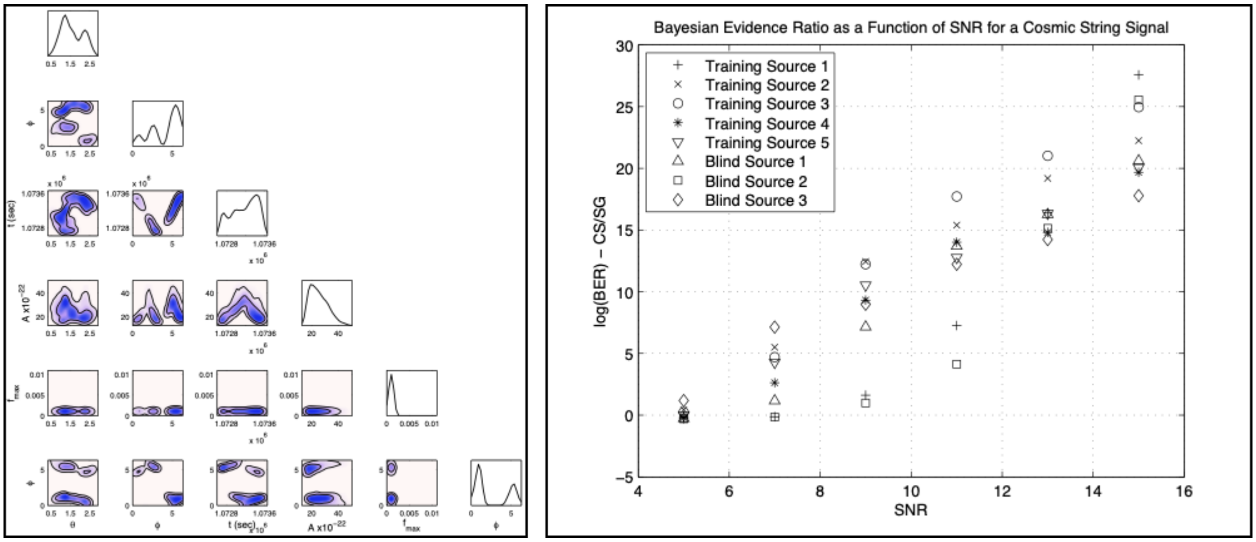


Figure 41: Left panel: posterior on the parameters characterising one of the cosmic string cusp gravitational wave bursts in the MLDC round 3 cosmic string data set. Right panel: evidence ratio in favour of the true (cosmic string cusp) model versus an alternative (sine-Gaussian) model for the burst, as a function of the burst signal-to-noise ratio. Figures reproduced from Feroz, F., Gair, J.R., Graff, P., Hobson, M.P., and Lasenby, A., *Class. Quantum Grav.* **27**, 075010 (2010).

J.R., Hobson, M.P., and Porter, E.K., *Class. Quantum Grav.* **26**, 215003). MultiNest was also used to find and characterise supermassive black hole mergers and gravitational wave bursts from cosmic string cusps in MLDC data. In the latter case, the computed Bayesian evidences were used to test the hypothesis that the burst signals were consistent with a cosmic string cusp as opposed to a generic sine-Gaussian burst model (see Figure **??** and Feroz, F., Gair, J.R., Graff, P., Hobson, M.P., and Lasenby, A., *Class. Quantum Grav.* **27**, 075010 (2010)).

Further details on LISA data analysis can be found in the MLDC papers, and references therein:

- Arnaud, K.A., et al. *The Mock LISA Data Challenges: An overview*, *AIP Conf. Proc.* **873**, 619 (2006).
- Arnaud, K.A., et al., A How-To for the Mock LISA Data Challenges, *AIP Conf. Proc.* **873**, 625 (2006).
- Arnaud, K.A., et al., *Report on the first round of the Mock LISA Data Challenges*, *Class. Quantum Grav.* **24**, S529 (2007).
- Arnaud, K.A., et al., *An overview of the second round of the Mock LISA Data Challenges*, *Class. Quantum Grav.* **24**, S551 (2007).
- Babak, S., et al., *Report on the second Mock LISA Data Challenge*, *Class. Quantum Grav.* **25**, 114037 (2008).
- Babak, S., et al., *The Mock LISA Data Challenges: from Challenge 1B to Challenge 3*, *Class. Quantum Grav.* **25**, 184026 (2008).

- Babak, S., et al., *The Mock LISA Data Challenges: from Challenge 3 to Challenge 4*, *Class. Quantum Grav.* **27**, 084009 (2010).

9 Time Series

We encountered the notion of a time series, or stochastic process, in Section 6 when we discussed modelling of the noise in gravitational wave detectors. In this section we will described some more general properties of time series, and several families of time series that might be encountered when analysing data. The basic idea of a time series if that it is an ordered sequence of random variables, such that each subsequent value depends on (in the sense of being correlated with) previous values. There are two main types of time series

- Available data are part of a **random sequence** $\{X_t\}$, which is only defined at integer values of the time t.
- Available data are values of a **random function**, $X(t)$, that is defined for arbitrary $t \in \mathbb{R}$, but is only observed at a finite number of times.

Random functions can be represented as random sequences, e.g., by integrating or averaging, but in general this throws away information, so where possible it is better to treat the function as continuous when performing an analysis.

We conclude this preamble with some definitions. Let $\{X_t\}_{t\in\mathcal{T}}$ be a stochastic process, then

1. if $\mathbb{E}(X_t) < \infty$, then the **mean** (or **expectation**) of the process is
$$\mu_t = \mathbb{E}(X_t).$$
If μ_t is non-constant, i.e., it depends on t, then μ_t is sometimes called the **trend**.
2. if $\mathrm{var}(X_t) < \infty$ for all $t \in \mathcal{T}$, then the **(auto)covariance** function of the random process is defined as
$$\gamma(s,t) = \mathrm{cov}(X_s, X_t) = \mathbb{E}\left\{(X_s - \mu_s)(X_t - \mu_t)\right\}, \quad s,\, t \,\in \mathcal{T}$$
and the **(auto)correlation function** of the process is defined by
$$\rho(s,t) = \frac{\gamma(s,t)}{\{\gamma(s,s)\gamma(t,t)\}^{1/2}}, \quad s,\, t \,\in \mathcal{T}.$$
Note that $\mathrm{var}(X_t)$ =$\mathrm{cov}(X_t, X_t) = \gamma(t,t)$ and $|\rho(s,t)| \leq 1$ for all $s,\, t \,\in \mathcal{T}$ from the Cauchy-Schwarz inequality. In addition, the function $\gamma(s,t)$ is semi-positive definite, i.e.,
$$\sum a_i a_j \gamma(t_i, t_j) \geq 0$$
for any $\{a_1, \ldots, a_k\} \in \mathbb{R}$ and any $\{t_1, \ldots, t_k\}$.

9.1 General properties of time series

9.1.1 Stationarity

If $\mathcal{S}$ is a set, then we use $u + \mathcal{S}$ to denote the set $\{u + s : s \in \mathcal{S}\}$, and $X_{\mathcal{S}}$ to denote the set of random variables $\{X_s : s \in \mathcal{S}$. A stochastic process is said to be

- **strictly stationary** if for any finite subset $\mathcal{S} \subset \mathcal{T}$ and any u such that $u + \mathcal{S} \subset \mathcal{T}$, the joint distributions of $X_{\mathcal{S}}$ and $X_{\mathcal{S}+u}$ are the same;

- **second-order stationary** (or **weakly stationary**) if the mean is constant and the covariance function $\gamma(s,t)$ depends only on $|s-t|$.

When $\mathcal{T} = \mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$ and the process is stationary

$$\gamma(t, t+h) = \gamma(0, h) = \gamma(0, -h) \equiv \gamma_{|h|} = \gamma_h, \quad h \in \mathbb{Z},$$

where h is called the **lag**. Similarly $\rho(t, t+h) \equiv \rho_{|h|} = \rho_h$ for $h \in \mathbb{Z}$. So, in the stationary case the covariance and correlation functions are symmetric around $h = 0$.

In practice, it is impossible to verify strict stationarity and many computations require only second-order stationarity. Elsewhere in this chapter when we refer to “stationarity” we will mean second-order stationarity. Third and higher-order stationarity is defined analogously, by extending the definition to third or higher correlation moments. In cases where there is a trend or seasonality in the data, the time series will often be preprocessed to remove the trend and leave a stationary stochastic process that can be analysed using methods that assume stationarity. One way to do this is to use **differencing**.

9.1.2 Examples of stochastic processes

1. A stochastic process is called **white noise** if its elements are uncorrelated, $\mathbb{E}(X_t) = 0$ and variance $\mathrm{var}(X_t) = \sigma^2$. If the elements are normally distributed then it is a **Gaussian white noise** process, $X_t \sim^{\mathrm{iid}} N(0, \sigma^2)$. As all elements of the series are independent, this is clearly a stationary stochastic process.

2. A **random walk** is defined by

$$X_t = X_{t-1} + w_t, \qquad t = 1, 2, \ldots.$$

The expectation value of this process is 0, and the autocorrelation is $\gamma_h = 1$ for all h. However, it is not a stationary process because $\mathrm{var}(X_t)$ is infinite.

9.1.3 Differencing

We define the **backshift operator** B by $BX_t = X_{t-1}$ and the **first difference** of the series $\{X_t\}$ by $\{\nabla X_t\}$, where

$$\nabla X_t = (I - B)X_t = X_t - X_{t-1}$$

and **higher-order differences**, such as the second difference $\{\nabla^2 X_t\}$ by

$$\nabla^2 X_t = \nabla(\nabla X_t) = \nabla(X_t - X_{t-1}) = X_t - 2X_{t-1} + X_{t-2}$$

and so on. If $X_t = p(t) + w_t$, where $p(t)$ is a polynomial of degree k and $\{w_t\}$ is a stationary stochastic process, then $\{\nabla^k X_t\}$ is stationary, i.e., k'th order differencing removes the polynomial trend. For example, first-order differencing reduces a random walk to a stationary process. This procedure will be exploited when discussing ARIMA processes later in this chapter. When dealing with observed time-series, it is normal to apply successive differences to the data until the resulting time series appears to be stationary.

9.1.4 Causal processes

Suppose that the process $\{X_t\}$ can be written in the linear form

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}$$

where $\{w_t\}$ is white noise, $\sum |\psi_j| < \infty$, and $\psi_0 = 1$. The process is called **causal** if $\psi_{-1} = \psi_{-2} = \cdots = 0$, so the linear expression for X_t does not involve the future values of w_t.

Using the backshift operator B we can write $w_{t-j} = B^j w_t$, so

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j B^j w_t = \psi(B) w_t,$$

where

$$\psi(u) = \sum_{j=-\infty}^{\infty} \psi_j u^j$$

is an infinite series and $\psi(B)$ the corresponding operator. The properties of the polynomial defined here are crucial for determining properties of stationary time series such as invertibility, as we will see in the following sections.

9.2 Moving-average (MA) processes

One of the most commonly encountered types of stationary stochastic process is a moving average process. Let $\{w_t\} \sim (0, \sigma^2)$ be a white noise process for $t \in \mathbb{Z}$. Then the time series $\{X_t\}$ is said to be a **moving average process of order** q (denoted $\mathbf{MA}(q)$) if

$$X_t = w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}$$

where $\theta_1, \ldots, \theta_q$ are real valued constants.

The mean of X_t is

$$\begin{aligned} \mathbb{E}[X_t] &= \mathbb{E}[w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}] \\ &= \mathbb{E}[w_t] + \theta_1 \mathbb{E}[w_{t-1}] + \cdots + \theta_q \mathbb{E}[w_{t-q}] = 0. \end{aligned} \tag{111}$$

Setting $\theta_0 = 1$ the autocovariance is

$$\begin{aligned} \gamma(k) &= \mathrm{cov}(X_t, X_{t+k}) = \mathbb{E}[X_t X_{t+k}] - 0^2 \\ &= \mathbb{E}[(\theta_0 w_t + \cdots + \theta_q w_{t-q})(\theta_0 w_{t+k} + \cdots + \theta_q w_{t+k-q})] \\ &= \sum_{r=0}^{q} \sum_{s=0}^{q} \theta_r \theta_s \mathbb{E}[w_{t-r} w_{t+k-s}]. \end{aligned} \tag{112}$$

This can be simplified by noting

$$\mathbb{E}[w_{t-s} w_{t+k-r}] = \begin{cases} \sigma^2 & \text{if } t - r = t + k - s \\ 0 & \text{otherwise (since } w_t \text{ are uncorrelated).} \end{cases}$$

When $r, s \leq q$ then $t - r \neq t + k - s$ for any r, s if $|k| > q$ and so

$$\gamma(k) = \begin{cases} 0 & \text{if } |k| > q \\ \sigma^2 \sum_{r=0}^{q-|k|} \theta_r \theta_{r+|k|} & \text{if } |k| \leq q. \end{cases}$$

Since the mean is constant and $\gamma(k)$ does not depend on t, we see that MA(q) is a stationary stochastic process. The variance is

$$\mathrm{var}(X_t) = \gamma_0 = \sigma^2 \sum_{r=0}^{q} \theta_r^2$$

and the autocorrelation function is

$$\rho(k) = \begin{cases} 0 & \text{if } |k| > q \\ \sum_{r=0}^{q-|k|} \theta_r \theta_{r+|k|} / \sum_{r=0}^{q} \theta_r^2 & \text{if } |k| \leq q. \end{cases}$$

Note that $\rho(k) = 0$ for $|k| > q$. This fact is useful when detecting MA(q) processes in observed data.

The moving average process is a weighted sum of a finite number of white noise events. Applications within economics include modelling the effects of strikes on economic output (the white noise events are the strikes, but the impact on economic output at any given time is not only due to any current strikes, but also previous strikes), or modelling the sales of white goods (people replace white goods when they break, and those breakages are the white noise processes, but people might not all replace immediately, so there will be some influence of lags).

The autocorrelation function does not convey all information about a moving average process, since two different moving average processes may have the same auto-correlation function. This is most easily seen by an example. Consider the two processes

$$X_t = w_t + \theta w_{t-1} \quad \text{and} \quad X_t = w_t + \frac{1}{\theta} w_{t-1}.$$

The autocorrelation function of both of these processes is

$$\rho(1) = \rho(-1) = \frac{\theta}{1+\theta^2}, \qquad \rho(k) = 0 \quad \text{for } |k| > 1.$$

However, we can rearrange the first process to give

$$w_t = X_t - \theta X_{t-1} + \theta^2 X_{t-2} - \cdots$$

while rearranging the second process we obtain

$$w_t = X_t - \frac{1}{\theta} X_{t-1} + \frac{1}{\theta^2} X_{t-2} - \cdots .$$

If $|\theta| < 1$ the series of coefficients converges for the first model and not the second, and vice versa for $|\theta > 1$. This ambiguity leads to the notion of **invertibility**.

9.2.1 Invertible moving average processes

A general MA(q) process $\{X_t\}$ is said to be **invertible** if it can be written as a convergent sum of present and past values of X_t of the form

$$w_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$$

where $\sum |\pi_j| < \infty$. There is only one invertible MA(q) process associated with each autocorrelation function $\rho(k)$ and so this notion eliminates the ambiguity identified in the previous example. To determine if a MA(q) process is invertible we can use the backshift operator introduced above to write

$$\begin{aligned} X_t &= w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q} \\ &= (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q) w_t \\ &= \theta(B) w_t \end{aligned} \tag{113}$$

where $\theta(B)$ is the polynomial

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q.$$

Although this polynomial defines an operator, it can be manipulated in the same way as standard polynomials. In this way, it can be seen that the process is **invertible** if the roots of $\theta(B)$ all lie **outside the unit circle**, i.e., all (possibly complex) solutions to $\theta(z) = 0$ have $|z| > 1$.

Example: The MA(1) model $X_t = w_t + \theta_1 w_{t-1}$ can be written as

$$X_t = (1 + \theta_1 B) w_t \quad \Rightarrow \quad \theta(B) = 1 + \theta_1 B$$

which has a single root at $B = -1/\theta_1$. Therefore the process is invertible if $|\theta_1| < 1$.

9.3 Autoregressive (AR) processes

Another commonly encountered type of stationary stochastic process is an auto-regressive process. Let $w_t \sim (0, \sigma^2)$ for $t \in \mathbb{Z}$ as in the previous section. The time series $\{X_t\}$ is said to be an **autoregressive process of order** p (denoted **AR(**p**)**) if

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + w_t$$

where $\alpha_1, \alpha_2, \ldots, \alpha_p$ are constants. Autoregressive models assume current values of a time series depend on a fixed number of previous values (plus some random noise). An example from forensic science is the concentration of cocaine on bank notes in a bundle. Cocaine transfers between the notes and therefore there will be a correlation between consecutive notes in the bundle (ordering of the notes in the bundle is a proxy for time in this example).

Example: The autoregressive process of order one is

$$X_t = \alpha_1 X_{t-1} + w_t$$

which is closely related to the random walk process defined earlier. Through repeated substitution we see

$$X_t = \alpha_1(\alpha_1 X_{t-2} + w_{t-1}) + w_t = w_t + \alpha_1 w_{t-1} + \alpha_1^2 w_{t-2} + \cdots$$

so an AR(1) process can be written as in infinite order moving average process. The mean is

$$\mathbb{E}[X_t] = \mathbb{E}[w_t + \alpha_1 w_{t-1} + \alpha_1^2 w_{t-2} + \cdots] = 0$$

and the autocovariance function is

$$\begin{aligned}\gamma(k) = \mathrm{cov}(X_t, X_{t+k}) &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \alpha_1^i w_{t-i}\right)\left(\sum_{j=0}^{\infty} \alpha_1^j w_{t+k-j}\right)\right] \\ &= \sigma^2 \sum_{i=0}^{\infty} \alpha_1^i \alpha_1^{k+i} \text{ for } k \geq 0 \text{ since } \mathbb{E}[w_{t-i} w_{t+k-j}] = 0 \text{ unless } j = k+i \\ &= \frac{\sigma^2 \alpha_1^k}{(1-\alpha_1^2)} \text{ if } |\alpha_1| < 1. \qquad (114)\end{aligned}$$

Hence an AR(1) process with $|\alpha_1| < 1$ is stationary, with $\mathrm{var}(X_t) = \gamma(0) = \sigma^2/(1-\alpha_1^2)$ and autocorrelation $\rho(k) = \gamma(k)/\gamma(0) = \alpha_1^{|k|}$.

For the general AR(p) process, we can write

$$\begin{aligned}X_t - \alpha_1 X_{t-1} - \alpha_2 X_{t-2} - \ldots - \alpha_p X_{t-p} &= w_t \\ (1 - \alpha_1 B - \alpha_2 B^2 - \ldots - \alpha_p B^p) X_t &= w_t \\ \phi(B) X_t &= w_t. \qquad (115)\end{aligned}$$

Recall that a time series is causal if there exists $\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \ldots$ such that $\sum_{i=0}^{\infty} |\psi_i| < \infty$ and $X_t = \psi(B) w_t$. From the above result, any such $\psi(B)$ must be the inverse of $\phi(B)$. We deduce that the AR(p) process is causal if and only if all of the roots of the polynomial $\phi(u)$ **lie outside the unit circle**. If this is true, then the coefficients ψ_i can be found from the expansion of the function $1/\phi(B)$ in the usual way.

The mean and covariance of a causal AR(p) process can be found from the decomposition $X_t = \sum \psi_i w_{t-i}$ The mean is clearly zero and the covariance can be found from

$$\begin{aligned}\gamma(k) &= \mathrm{cov}(X_t, X_{t+k}) \\ &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \psi_i w_{t-i}\right)\left(\sum_{j=0}^{\infty} \psi_j w_{t+k-j}\right)\right] \\ &= \sigma^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k} \text{ for } k \geq 0. \qquad (116)\end{aligned}$$

The auto-covariance function converges (and hence $\{X_t\}$ is weakly stationary) if $\sum |\psi_i|$ converges, which was the condition for the series to be causal. So an AR(p) process is weakly stationary if it is causal.

Example: consider the AR(1) process

$$X_t = \alpha_1 X_{t-1} + w_t.$$

This may be written

$$\phi(B)X_t = w_t, \qquad \text{where } \phi(B) = (1 - \alpha_1 B).$$

The root of $\phi(B)$ is $B = 1/\alpha_1$, which lies outside the unit circle if $|\alpha_1| < 1$. Therefore, AR(1) models are causal (and weakly stationary) if $|\alpha_1| < 1$. If this is true then we can write

$$\begin{aligned} X_t &= \frac{1}{\phi(B)} w_t \\ &= (1 - \alpha_1 B)^{-1} w_t \\ &= (1 + \alpha_1 B + (\alpha_1 B)^2 + \ldots) w_t \\ &= \psi_0 w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} + \ldots \end{aligned} \tag{117}$$

with $\psi_i = \alpha_1^i$ for $i \in \{0, 1, 2, \ldots\}$. This agrees with the result obtained previously by repeated substitution of the original equation.

9.4 Estimating properties of stationary time series

9.4.1 Estimation

Suppose we have observed values $x_1, \ldots, x_n$ of a time series $\{X_t\}$ at times $t = 1, 2, \ldots, n$. We suppose that $\{X_t\}$ is weakly stationary so that $\mathbb{E}[X_t] = \mu$, $\gamma(k)$ and $\rho(k)$ exist. These three quantities can be estimated as follows

- We estimate μ by the sample mean

$$\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_i.$$

- We estimate $\gamma(k)$ at lag k by

$$c_k = \frac{1}{n-k-1} \sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x}).$$

  The estimator c_k is called the **sample autocovariance coefficient at lag** k.

- We estimate $\rho(k)$ at lag k by

$$r_k = \frac{c_k}{c_0},$$

  and this estimator is referred to as the **sample autocorrelation coefficient at lag** k. A plot of r_k against k is called a **correlogram**.

The latter two formulas are only valid if k is small relative to n, roughly $k < n/3$.

9.4.2 Tests for a white noise process

If $\{X_t\}$ is a white noise process (plus possibly a constant mean), then for large n

$$r_k \dot{\sim} N(0, 1/n).$$

To test the hypothesis H_0 that the process $\{X_t\}$ is white noise we can use the values of the r_k's. Rather than treating each r_+k as an independent test statistic, it is better to count the number of r_k's that exceed a relevant threshold. For example, for a 5% significance test we compare each $|r_k|$ to $1.96/\sqrt{n}$ and count the number, b say, that exceed this value. Under H_0

$$b \dot{\sim} \mathrm{Bin}(m, 0.05)$$

where m is the number of r_k's being computed. Roughly speaking, if b exceeds $m/20$ then we would reject H_0.

Another test for white noise is the **portmanteau test** (Box and Pierce 1970; Ljung and Box 1978). If $m \ll n$ and $n \gg 1$, then

$$Q_m = n(n+2) \sum_{h=1}^{m} (n-h)^{-1} \hat{\rho}_h^2 \dot{\sim} \chi_m.$$

The sensitivity of Q_m to different types of departure from white noise depends on m. If m is too large, sensitivity is reduced because some of the $\hat{\rho}_h$ will contribute no information about the lack of fit. If m is too small then sensitivity is reduced because some of the $\hat{\rho}_h$ that convey information about the lack of fit are missing.

9.4.3 Testing for stationarity

One common test for stationarity is based on fitting the model

$$X_t = \xi t + \eta_t + \epsilon_t, \qquad \eta_t = \eta_{t-1} + w_t, \qquad w_t \sim^{\text{iid}} (0, \sigma_w^2)$$

where $\{\epsilon_t\}$ is assumed to be stationary. If $\sigma_w^2 > 0$ then the sequence is a random walk. If $\sigma_w = 0$ and $\xi = 0$ then the series is called **level stationary** since $\{X_t\}$ is stationary. If $\sigma_w = 0$ but $\xi \neq 0$ it is called **trend stationary** as then $\{X_t - \xi t\}$ is stationary.

The **KPSS** test for stationarity is based on a score test for the hypothesis that $\sigma_w^2 = 0$, leading to

$$C(l) = \hat{\sigma}(l)^{-2} \sum_{t=1}^{n} S_t^2, \qquad \text{where } S_t = \sum_{j=1}^{t} e_j, \qquad t = 1, \ldots, n,$$

where $e_1, \ldots, e_n$ are the residuals from a straight-line regression to the data, $X_t = \alpha + \beta t + \epsilon_t$, and $\hat{\sigma}(l)^2$ is the estimated variance based on residuals truncated at lag l. Under certain assumptions, $C(l)$ has a tractable asymptotic distribution (integral of a squared Brownian bridge).

9.4.4 Detection of MA(q) processes

As discussed earlier, $\rho(k) = 0$ for $|k| > q$ for an MA(q) process. Hence if $\{X_t\}$ are from a MA(q) process, we would expect

1. $r_1, r_2, \ldots, r_q$ will be fairly close to $\rho(1), \rho(2), \ldots, \rho(q)$ (and hence not close to 0).

2. $r_{q+1}, r_{q+2}, \ldots$ will be randomly distributed about zero.

Inspection of the sample autocorrelation coefficients can thus identify moving average processes. For example, if $|r_1|$ was large but $r_2, r_3, \ldots$ are close to zero, there would be evidence that is was a MA(1) process.

9.4.5 Detection of AR(p) processes

In an AR(1) process $X_t = \alpha_1 X_{t-1} + w_t$, the autocorrelation function is given by

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \alpha_1^{|k|}.$$

Therefore, the sample autocorrelation coefficient, r_1, gives an estimate of α_1, and the other sample autocorrelation coefficients should scale like $r_1^{|k|}$. Note that, unlike the MA(q) model, the coefficients, r_k, do not drop to zero above some threshold.

For a general AR(p) process, detecting the order of the process by inspection of the coefficients is difficult. Instead, to fit the general AR(p) model

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + w_t$$

we can find the coefficients that minimize

$$\frac{1}{n} \sum_{t=p+1}^{n} \left(x_t - \sum_{i=1}^{p} \alpha_i x_{t-i} \right)^2 .$$

The resulting estimates $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_p$ are known as least squares estimates for obvious reasons. The estimate $\hat{\alpha}_p$ is also called the **sample partial autocorrelation coefficient at lag** p. This provides an estimate of the the autocorrelation at lag p that is not accounted for by the autocorrelation at smaller lags, hence the term "partial". A plot of the sample partial autocorrelation coefficients versus lag is called the **partial autocorrelation function (pacf)** and is analogous to the correlogram. For an AR(p) process, the partial autocorrelation coefficients $\hat{\alpha}_{p+1}, \hat{\alpha}_{p+2}, \ldots$ should drop to around zero. Hence, they can be used to estimate the order of an AR process in the same way that the correlogram can be used to estimate the order of a MA process. The partial autocorrelation coefficient at lag k is significantly different from zero at the 5% significance level if it is outside the range $(-2/\sqrt{n}, 2/\sqrt{n})$.

9.4.6 Time series residuals

The **residuals** of a time series are defined as

$$\hat{w}_t = \text{ observation } - \text{ fitted value.}$$

For example, for an AR(1) model, $X_t = \alpha X_{t-1} + w_t$, with observations $\{x_t\}$, $t \in \{1, 2, \ldots, n\}$, the residuals are given by

$$\hat{w}_t = x_t - \hat{\alpha} x_{t-1},$$

where $\hat{\alpha}$ is the estimate of the parameter α, obtained for example from the least squares estimation procedure described above. The fitted value at time t is the forecast of x_t, made at time $t-1$.

For a model that fits well, the residuals $\{w_t\}$ will be approximately white noise, with constant variance. There are three standard approaches to assessing time series residuals

1. Plotting the residuals versus time. The residuals should be uncorrelated and randomly distributed about zero. Any patterns in the data, or significant outliers suggest that the model is not well fitted.

2. Use the Ljung-Box statistic defined above.

3. Looking at the correlogram of the residuals. Any autocorrelation coefficients lying outside the range $\pm 2/\sqrt{n}$ can be said to be significantly different from zero at the 5% significance level.

Note that the residuals are not exactly white noise, so these tests must not be used precisely, but are guidelines.

9.5 ARMA processes

An ARMA(p,q) process is a combination of an MA(q) and an AR(p) process. The time series $\{X_t\}$ is said to be an ARMA(p,q) process if X_t is given by

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}$$

where $w_t \sim (0,\sigma^2)$ is a white noise process as usual. Using the backshift operator we can write the ARMA(p,q) process as

$$\phi(B)X_t = \theta(B)w_t$$

where $\phi(B) = 1 - \alpha_1 B - \alpha_2 B^2 - \ldots - \alpha_p B^p$ and $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$. Moving average, autoregressive and white noise process are all special cases of ARMA models. An MA(q) process is an ARMA$(0,q)$ model, an AR(p) process is ARMA$(p,0)$ and white noise is an ARMA$(0,0)$ process.

It is useful for ARMA(p,q) models to be both causal and invertible and the conditions for this are the same as the conditions for invertibility of the MA(q) process and causality of the AR(p) process, namely

- For an ARMA(p,q) process to be **invertible**, the roots of $\theta(B)$ must lie outside the unit circle.

- For an ARMA(p,q) process to be **causal**, the roots of $\phi(B)$ must lie outside the unit circle.

If an ARMA(p,q) process is both invertible and causal then it can be expressed both as an infinite order moving average process and as an infinite order autoregressive process.

An ARMA(p,q) process is **regular** if

1. It is both invertible and causal,

2. $\theta(B)$ and $\phi(B)$ have no common roots.

The second condition is necessary because if the two functions have a common root, the process can be simplified to one with fewer terms.

If an ARMA(p, q) process is regular then it maybe written

$$X_t = \frac{\theta(B)}{\phi(B)} w_t = \psi(B) w_t$$

where

$$\psi(B) = \frac{\theta(B)}{\phi(B)} = \psi_0 + \psi_1 B + \psi_2 B^2 + \ldots = \sum_{i=0}^{\infty} \psi_i B^i$$

with $\psi_0 = 1$ and $\sum_{i=0}^{\infty} |\psi_i < \infty$. In other words

$$X_t = w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} + \ldots$$

This is an infinite order moving average process and is known as the **Wold decomposition** of X_t.

In the same way, it is also possible to express w_t in terms of X_t using

$$w_t = \frac{\phi(B)}{\theta(B)} X_t = \pi(B) X_t = \sum_{i=0}^{\infty} \pi_i X_{t-i}$$

where

$$\pi(B) = \frac{\phi(B)}{\theta(B)} = 1 + \pi_1 B + \pi_2 B^2 + \ldots = \sum_{i=0}^{\infty} \pi_i B^i$$

with $\pi_0 = 1$. This inversion formula is used in some **forecasting** methods.

For a regular ARMA(p, q) process we have

$$\rho(k) = \frac{\sum_{i=0}^{\infty} \psi_i \psi_{i+k}}{\sum_{i=0}^{\infty} \psi_i^2} \quad \text{for } k = 1, 2, \ldots.$$

This can be proved as follows. Firstly we note

$$\begin{aligned} \gamma(k) &= \text{cov}(X_t, X_{t+k}) = \mathbb{E}[X_t X_{t+k}] - 0 \\ &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \psi_i w_{t-i}\right)\left(\sum_{j=0}^{\infty} \psi_j w_{t+k-j}\right)\right] \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \psi_i \psi_j \mathbb{E}(w_{t-i} w_{t+k-j}). \end{aligned} \tag{118}$$

Now

$$\mathbb{E}[w_{t-i} w_{t+k-j}] = \begin{cases} \sigma^2 & \text{if } j = i + k \\ 0 & \text{otherwise (since } w_t \text{ are uncorrelated.} \end{cases}$$

Therefore

$$\gamma(k) = \sigma^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k}$$

and

$$\gamma(0) = \sigma^2 \sum_{i=0}^{\infty} \psi_i^2.$$

Taking the ratio $\rho(k) = \gamma(k)/\gamma(0)$ we deduce the result quoted above.

Example: Consider an ARMA(1,1) process defined by

$$X_t = \alpha X_{t-1} + w_t + \beta w_{t-1}$$

where $\alpha, \beta \neq 0$ and $\{w_t\}$ is a Gaussian white noise process. Using the previous notation we have

$$\phi(B) = (1 - \alpha B), \qquad \theta(B) = (1 + \beta B).$$

The process is regular if the roots of $\phi(B)$ and $\theta(B)$ lie outside the unit circle and there are no roots in common. This is satisfied if

$$|\alpha| < 1, \quad |\beta| < 1 \quad \text{and } \alpha \neq -\beta.$$

If we now assume that these conditions are satisfied so the process is regular, we can use the Wold decomposition to obtain the variance and auto-correlation function. First we note

$$\begin{aligned} X_t &= \frac{1 + \beta B}{1 - \alpha B} w_t \\ &= (1 + \alpha B + \alpha^2 B^2 + \ldots)(1 + \beta B) w_t \\ &= [(1 + \alpha B + \alpha^2 B^2 + \ldots) + (\beta B + \beta\alpha B^2 + \beta\alpha^2 B^3 + \ldots)] w_t \\ &= [1 + (\alpha + \beta)B + (\alpha^2 + \alpha\beta)B^2 + (\alpha^3 + \alpha^2\beta)B^3 + \ldots] w_t \\ &= \sum_{i=0}^{\infty} \psi_i w_{t-i} \end{aligned} \tag{119}$$

where $\psi_i = (\alpha + \beta)\alpha^{i-1}$ for $i = 1, 2, \ldots$ and $\psi_0 = 1$. Using this decomposition we can compute the variance

$$\begin{aligned} \text{var}[X_t] &= \sum_{i=0}^{\infty} \psi_i^2 \text{var}[w_{t-i}] = \sigma^2 \sum_{i=0}^{\infty} \psi_i^2 \\ &= [1 + (\alpha + \beta)^2 + (\alpha + \beta)^2\alpha^2 + (\alpha + \beta)^2\alpha^4 + \ldots]\sigma^2 \\ &= \left[1 + \frac{(\alpha + \beta)^2}{(1 - \alpha^2)}\right] \sigma^2. \end{aligned} \tag{120}$$

The autocorrelation function can be found from the formula

$$\rho(k) = \frac{\sum_{i=0}^{\infty} \psi_i \psi_{i+k}}{\sum_{i=0}^{\infty} \psi_i^2}.$$

For example, for $k = 1$, we have from the variance result

$$\sum_{i=0}^{\infty} \psi_i^2 = \left[1 + \frac{(\alpha + \beta)^2}{(1 - \alpha^2)}\right] = \frac{1 + 2\alpha\beta + \beta^2}{1 - \alpha^2}$$

and note

$$\begin{aligned} \psi_0\psi_1 + \psi_1\psi_2 + \psi_2\psi_3 + \ldots &= (\alpha + \beta) + [(\alpha + \beta)^2\alpha + (\alpha + \beta)^2\alpha^3 + \ldots] \\ &= (\alpha + \beta) + \left[\frac{(\alpha + \beta)^2\alpha}{1 - \alpha^2}\right]. \end{aligned} \tag{121}$$

Hence we find

$$\rho(1) = \frac{(\alpha + \beta)[(1 - \alpha^2) + (\alpha + \beta)\alpha]}{1 + 2\alpha\beta + \beta^2} = \frac{(\alpha + \beta)[1 + \alpha\beta]}{1 + 2\alpha\beta + \beta^2}.$$

9.5.1 ARMA(p, q) with constant mean

The ARMA(p, q) model can be generalised to

$$X_t = c + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}$$

or equivalently

$$\phi(B)X_t = c + \theta(B)w_t$$

where $c \neq 0$. This is called an **ARMA(p, q) model with constant mean**. By letting

$$\mu = \frac{c}{1 - \alpha_1 - \alpha_2 - \ldots - \alpha_p} = \mathbb{E}[X_t]$$

the problem may be converted to a model with no constant term by considering

$$Y_t = X_t - \mu.$$

We can see that

$$\begin{aligned} \phi(B)Y_t &= \phi(B)(X_t - \mu) = \phi(B)X_t - \phi(B)\mu \\ &= c + \theta(B)w_T - c = \theta(B)w_t \end{aligned} \tag{122}$$

so $Y_t \sim$ARMA(p, q). If the ARMA process is regular then

$$Y_t = \frac{\theta(B)}{\phi(B)} w_t = \psi(B)w_t$$

and $X_t = Y_t + \mu$, from which we deduce

$$X_t = \mu + \sum_{i=0}^{\infty} \psi_i w_{t-i}.$$

The autocorrelation function $\rho(k)$ is the same for X_t and Y_t, as it does not depend on the value of μ.

9.6 ARIMA processes

The ARMA(p, q) models describe stationary time series, but often an observed time series $\{X_t\}$ is not stationary. To fit a stationary model to the data it is necessary to first remove the non-stationary behaviour, for example if the trend, $\mathbb{E}[X_t]$, is not constant. One approach is to consider differences of the time series, as these will remove polynomial trends as discussed earlier.

We denote the backward difference operator, $(I - B)$, by ∇. If $\{X_t\}$ has a trend which follows a polynomial of degree $\leq d$ in time, t, then we consider the d-th order difference process

$$W_t = \nabla^d X_t = (I - B)^d X_t.$$

If the time series $\{W_t\}$ generated in this way can be modelled using an ARMA(p, q) process, then the series is called an **autoregressive integrated moving-averaged (ARIMA) model** and is denoted by ARIMA(p, d, q). The process $\{W_t\}$ may be a zero mean ARMA(p, q)

process, in which case the trend of the original series, $\mathbb{E}[X_t]$, is a polynomial of degree $\leq d-1$ and we may write

$$\phi(B)W_t = \theta(B)w_t.$$

Alternatively, the process $\{W_t\}$ may have a constant mean, in which case $\mathbb{E}[X_t]$ is a polynomial of degree d and we may write

$$\phi(B)W_t = c + \theta(B)w_t \text{ with } c \neq 0.$$

If the ARMA(p, q) process that models $\{W_t\}$ is regular then the polynomials $\phi(B)$ and $\theta(B)$ have no roots outside the unit circle. Writing

$$\Phi(B) = \phi(B)(I - B)^d$$

we have

$$\Phi(B)X_t = \phi(B)(I - B)^d X_t = \phi(B)W_t = \theta(B)w_t.$$

The process $\{X_t\}$ is invertible since the roots of $\theta(B)$ lie outside the unit circle and so we may write

$$w_t = \frac{\Phi(B)}{\theta(B)}X_t = \Pi(B)X_t = X_t + \pi_1 X_{t-1} + \pi_2 X_{t-2} + \ldots.$$

In addition we note that

$$1 + \pi_1 + \pi_2 + \ldots = 0.$$

This follows from the fact that

$$\Pi(B)\theta(B) = \Phi(B) = \phi(B)(I - B)^d \quad \Rightarrow \quad \Pi(1)\theta(1) = 0 \quad \Rightarrow \quad \Pi(1) = 0.$$

The last step follows from the fact that $\theta(1) \neq 0$ since by assumption alll of the roots of $\theta(B)$ lie outside the unit circle. While ARIMA(p, q) processes are invertible, they are not causal, since $(I - B)^d$ has d roots on the unit circle and hence so does $\Phi(B)$. Thus the Wold decomposition cannot be used for ARIMA processes.

Example: Consider the model

$$X_t = X_{t-1} + w_t - \theta w_{t-1}, \quad \text{with } 0 < |\theta| < 1 \text{ and } \mathbb{E}[X_t] = \mu.$$

We can write

$$W_t = X_t - X_{t-1} = w_t - \theta w_{t-1}$$

so $W_t \sim$ARMA$(0, 1)$ and hence $X_t \sim$ARIMA$(0, 1, 1)$. We have

$$\Phi(B)X_t = \theta(B)w_t, \qquad \text{where } \Phi(B) = (I - B), \quad \theta(B) = I - \theta B.$$

We can invert this process to obtain

$$\begin{aligned} w_t = \Pi(B)X_t &= \frac{I - B}{I - \theta B}X_t \\ &= (1 - B)(1 + \theta B + \theta^2 B^2 + \ldots)X_t \\ &= [1 - (1 - \theta)B - (1 - \theta)\theta B^2 - (1 - \theta)\theta^2 B^3 + \ldots]X_t \\ &= \sum_{i=0}^{\infty} \pi_i X_{t-i} \end{aligned} \tag{123}$$

where $\pi_i = -(1-\theta)\theta^{i-1}$. We can also confirm

$$\sum_{i=1}^{\infty} \pi_i = -(1-\theta)\sum_{i=0}^{\infty} \theta^i = -(1-\theta)\frac{1}{1-\theta} = -1 \quad \Rightarrow \quad 1 + \sum_{i=1}^{\infty} \pi_i = 0.$$

9.6.1 ARIMA processes with a constant term

Suppose that we have

$$\phi(B)(I-B)^d X_t = c + \theta(B)w_t,$$

where $c \neq 0$. This means that $\{X_t\}$ has a trend term which is a polynomial of degree d. To work with such a series we define a new series, $\{Y_t\}$, as

$$Y_t = X_t - At^d, \quad \text{where } A = \frac{c}{d!(1-\alpha_1-\alpha_2-\ldots-\alpha_p)}.$$

The new series is an ARIMA model without a constant term

$$\phi(B)(I-B)^d Y_t = \theta(B)w_t$$

and so can be used for forecasting. Forecasts of X_t can be obtained by adding At^d to the forecasts of Y_t.

10 Nonparametric Regression

The notes in this section are taken from a lecture course on this topic that I gave previously. We will not cover all of this material in one lecture, but the detailed notes are provided so that you can learn about more about the topics that interest you.

10.1 Introduction

10.1.1 Difference between parametric and nonparametric regression

The basis for regression is a set of observations of pairs of variables (X_i, Y_i), $i = 1, \ldots, n$. We are interested in finding a connection between X and Y. We assume that Y is random, but X can be either random or fixed; we focus mostly on the case that the X_i's are fixed. In parametric regression we assume a particular type of dependence of Y on X (e.g. linear regression: $\mathbb{E}Y = AX$, log-linear regression $\log(\mathbb{E}Y) = AX$, etc). In other words, we assume a priori that the unknown regression function f belongs to a parametric family $\{g(x, \theta) : \theta \in \Theta\}$, where $g(\cdot, \cdot)$ is a given function, and $\Theta \subset \mathbb{R}^k$. Estimation of f is the equivalent to estimation of the parameter vector θ.

In nonparametric regression, by contrast, we do not want to make any assumption about how $\mathbb{E}Y$ depends on X, but want to fit an arbitrary functional dependence. We assume that we observe a function with error:

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$

Often the errors are assumed to be normally distributed, $\varepsilon_i \sim N(0, \sigma^2)$, independently. The aim is to estimate the unknown function f.

In nonparametric estimation it is usually assumed that f belongs to some large class $\mathcal{F}$ of functions. For example, $\mathcal{F}$ can be the set of all the continuous functions or the set of all smooth (differentiable) functions. For proving certain properties of estimators, we will consider sets of functions with k derivatives, which are called Hölder spaces of functions.

We will described several different approaches to nonparametric regression — kernel smoothing, spline smoothing, general additive models and wavelet estimation.

10.1.2 Nonparametric regression model

Throughout this chapter we will assume the following model of nonparametric regression:

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$

with independent errors $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ and a function $f : [0, 1] \to \mathbb{R}$.

Now suppose that we observe data (x_i, y_i), $i = 1, \ldots, n$, which is a realisation of iid random variables (X_i, Y_i). The aim is to estimate the unknown function $f(x) = \mathbb{E}(Y_i | X_i = x)$, namely to construct an estimator $\widehat{f_n}(x)$ for all $x \in [0, 1]$ which is consistent and efficient, and to be able to test hypotheses about $f(x_0)$ for a fixed x_0 and about $f(x)$ for all x simultaneously.

The maximum likelihood estimator (MLE) of $f(x)$ gives estimates of f only at points x_i where we observe the data: $\widehat{f}(x_i) = y_i$. Since $\mathbb{E}[\varepsilon_i] = 0$, this estimator is unbiased at x_i, as $\mathbb{E}\widehat{f}(x_i) = \mathbb{E}Y_i = f(x_i)$. However, the MLE (and the model) does not give any information about $f(x)$ for $x \neq x_i$. The model is not fully identifiable hence some additional assumptions about f are needed. A key assumption we will make about f that it is smooth.

10.1.3 Estimators

There are two major approaches to nonparametric estimation.

1. **Smoothing**: fitting a flexible smooth curve to data. We will consider two methods: kernel smoothing and spline smoothing. The main question in this context is how smooth should this curve be, and do we have to decide that in advance, or can we let the data to decide?

2. **Orthogonal projection estimation**: represent the regression function f as a series in an orthogonal basis, and estimate the coefficients from the data. We will consider wavelet bases. Wavelets can be spiky, so they are well suited for modelling not very smooth functions, e.g., with jumps or sharp spikes. The main question is how to estimate the coefficients, so that the function estimate is neither too smooth nor too spiky.

10.1.4 Consistency

The key requirement for any estimator is consistency, that is, the more data we have, the closer the estimator is to the function of interest. We encountered consistency in the context of estimators of parameters, and there is a corresponding definition for functions.

Definition 10.1. *$\widehat{f}_n$ is a (weakly) consistent estimator of f in distance d based on n observations iff*

$$\forall \epsilon > 0, \quad \mathbb{P}(d(\widehat{f}_n, f) > \epsilon) \to 0 \quad \text{as} \quad n \to \infty.$$

In the rest of this chapter, when we refer to consistency we will mean weak consistency. We consider two distances on function spaces $d(\widehat{f}_n, f)$.

1) Pointwise at x_0 (local): $d(\widehat{f}_n, f) = |\widehat{f}_n(x_0) - f(x_0)|$, for some $x_0 \in [0, 1]$.

2) Integrated (global) : $d(\widehat{f}_n, f) = ||\widehat{f}_n - f||_2 = \sqrt{\int_0^1 (\widehat{f}_n(x) - f(x))^2 dx}$.

   Here $|| \cdot ||_2$ is defined by

$$||g||_2^2 \overset{def}{=} \int_0^1 [g(x)]^2 dx.$$

   It is a norm in Hilbert space $L^2[0, 1] = \{g : [0, 1] \to \mathbb{R} \text{ such that } ||g||_2 < \infty\}$.

Markov's inequality is a tool to verify consistency:

$$\mathbb{P}(d(\widehat{f}_n, f) > \epsilon) \leq \epsilon^{-2} \mathbb{E}[d(\widehat{f}_n, f)^2].$$

For these distances, $\mathbb{E}[d(\widehat{f}_n, f)]^2$ has particular names.

1) Mean squared error (MSE):

$$\text{MSE}(\widehat{f}_n(x_0)) = \mathbb{E}[|\widehat{f}_n(x_0) - f(x_0)|^2] = v(x_0) + [b(x_0)]^2$$

2) Mean integrated squared error (MISE):

$$\text{MISE}(\widehat{f}_n) = \mathbb{E}[||\widehat{f}_n - f||^2] = \mathbb{E}\left[\int_0^1 |\widehat{f}_n(x) - f(x)|^2 dx\right] = \int_0^1 v(x) dx + \int_0^1 [b(x)]^2 dx,$$

   where $b(x) = \text{bias}(\widehat{f}(x)) = \mathbb{E}\left[\widehat{f}(x)\right] - f(x)$ and $v(x) = \text{Var}(\widehat{f}(x))$ are the bias and the variance of $\widehat{f}(x)$.

Therefore, M(I)SE($\widehat{f}_n$) $\to 0$ as $n \to \infty$ implies consistency in the corresponding distance. We will also study the rate of convergence of the estimators, that is, how fast MISE and MSE decrease to 0 as a function of sample size n.

10.1.5 Notation

The indicator function of a set A is

$$\mathbf{1}_A(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{if } x \notin A. \end{cases}$$

Informally, we will also write $\mathbf{1}(|x| \leq 1)$ for $\mathbf{1}_{|x|\leq 1}(x)$.

Denote the support of a function g, the set of arguments where g is nonzero, by

$$\mathrm{supp}(g) = \{x : g(x) \neq 0\}.$$

10.2 Kernel estimators

10.2.1 Designs

Definition 10.2. *A set $(X_1, \ldots, X_n)$ is called a design*

Definition 10.3. *A design $(X_1, \ldots, X_n)$ is called fixed if the values $x_1, \ldots, x_n$ are non random*

Example 10.1. *An equispaced (regular) design $x_1 < x_2 < \ldots < x_n$ is a fixed design such that $x_i - x_{i-1} = 1/n$, e.g. $x_i = i/n$;$x_i = \frac{i-1}{n}$; $x_i = \frac{1}{2n} + \frac{i-1}{n}$.*

Definition 10.4. *A design $(X_1, \ldots, X_n)$ is called random iff $X_1, \ldots, X_n$ are iid random variables, $X_i \sim p(x)$.*

Example 10.2. *$x_i \sim U[0,1]$ with $p(x) = 1$ for $x \in [0,1]$.*

10.2.2 Nadaraya-Watson estimator

Definition 10.5. *A function $K(x)$ is called a kernel iff $\int_{-\infty}^{\infty} K(x)d(x) = 1$.*

If $K(x) \geq 0$, $K(x)$ is a probability density.

Definition 10.6. *If $K(x) = K(-x)$, then $K(x)$ is a symmetric kernel.*

Definition 10.7. *A kernel K has order m iff $\int_{-\infty}^{\infty} x^\ell K(x)dx = 0$ for all $\ell = 1, 2, \ldots, m-1$ and $\int_{-\infty}^{\infty} x^m K(x)dx \neq 0$.*

If K is symmetric, then K has order ≥ 2.

Example 10.3. *All these kernels are symmetric of order 2, except the last one.*

a) Uniform (box, rectangular) kernel $K(x) = \frac{1}{2}\mathbf{1}(|x| \leq 1)$.

b) Triangular kernel $K(x) = (1 - |x|)\mathbf{1}(|x| \leq 1)$.

c) Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

d) Cosine kernel $K(x) = \frac{\pi}{4}\cos(\pi x/2)\mathbf{1}(|x| \leq 1)$.

e) Sinc kernel $K(x) = \frac{\sin(\pi x)}{\pi x}$. *This kernel has infinite order, since* $\int_{-\infty}^{+\infty} \sin(\pi x)x^{m-1}dx = 0$ *for all integer* $m \geq 1$.

Remark 10.1. *If* $K(x)$ *is a kernel, then* $K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right)$ *is also a kernel. h is called the* **bandwidth**.

Example 10.4. *If* $K(x) = \frac{1}{2}\mathbf{1}(|x| \leq 1)$ *is a kernel then* $K(x) = \frac{1}{4}\mathbf{1}(|x| \leq 2)$ *is a kernel.*

Definition 10.8. ***The Nadaraya-Watson Estimator***

$$\widehat{f}_n^{NW}(x) = \frac{\sum_{i=1}^n Y_i K_h(X_i - x)}{\sum_{j=1}^n K_h(X_j - x)}, \text{ when } \sum_{i=1}^n K_h(X_i - x) \neq 0,$$

otherwise $\widehat{f}_n^{NW}(x) = 0$.

Motivation for the Nadaraya-Watson estimator.

Recall that $f(x)$ can be written as

$$f(x) = \mathbb{E}(Y_i \mid X_i = x) = \int yp(y \mid x)dy = \int \frac{yp(x,y)}{p(x)}dy.$$

Consider the following kernel density estimators:

$$\widehat{p}_n(x) = \frac{1}{n}\sum_{i=1}^n K_h(x_i - x), \quad \widehat{p}_n(x,y) = \frac{1}{n}\sum_{i=1}^n K_h(x_i - x)K_h(y_i - y). \tag{124}$$

Plugging $\widehat{p}_n(x)$ and $\widehat{p}_n(x,y)$ into $\mathbb{E}(Y_i|X_i = x)$, we have

$$\widehat{f}_h(x) = \int_{-\infty}^{\infty} \frac{y\widehat{p}_n(x,y)}{\widehat{p}_n(x)}dy.$$

Now we simplify the numerator, assuming that the kernel is symmetric

$$\int_{-\infty}^{\infty} y\widehat{p}_n(x,y)dy = \frac{1}{n}\int_{-\infty}^{\infty} y\sum_{i=1}^n K_h(x_i - x)K_h(y_i - y) = \frac{1}{n}\sum_{i=1}^n K_h(x_i - x)\int_{-\infty}^{\infty} yK_h(y - y_i)dy,$$

and the last integral is

$$\begin{aligned}\frac{1}{h}\int_{-\infty}^{\infty} yK\left(\frac{y - y_i}{h}\right)dy &= [z = (y - y_i)/h] = \int_{-\infty}^{\infty}(hz + y_i)K(z)dz \\ &= y_i\int_{-\infty}^{\infty} K(z)dz + h\int_{-\infty}^{\infty} zK(z)dz = y_i\end{aligned}$$

assuming that the order of the kernel K is at least 2.

Therefore, an estimator of f can be written as

$$\widehat{f}_h^{NW}(x) = \frac{n^{-1}\sum_{i=1}^n K_h(x_i - x)y_i}{n^{-1}\sum_{i=1}^n K_h(x_i - x)}\mathbf{1}\left(\sum_{i=1}^n K_h(x_i - x) \neq 0\right)$$

which coincides with the **Nadaraya-Watson estimator**. Thus, we proved the following proposition.

Proposition 10.1. *If $K(x)$ is a symmetric kernel of order ≥ 2, under random design,*

$$\widehat{f}_h^{NW}(x) = \int_{-\infty}^{\infty} \frac{y\widehat{p}_n(x,y)}{\widehat{p}_n(x)} dy\, \mathbf{1}(\widehat{p}_n(x) \neq 0),$$

where $\widehat{p}_n(x)$ and $\widehat{p}_n(x,y)$ are kernel density estimators defined by (124)*.*

If we know $p(x)$, then we can write $\widehat{f}(x) = \frac{1}{np(x)} \sum_{i=1}^n y_i K_h(x_i - x)$

If $X_i \sim U[0,1]$ then $p(x) = 1$ and $\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^n y_i K_h(x_i - x)$. This estimator also works for a regular fixed design.

Example 10.5. *Consider the box kernel $K(z) = 0.5\mathbf{1}(z \in [-1,1])$. Then, for x and h such that $|x_i - x| \leq h$ for at least one i, the Nadaraya-Watson estimator can be written as*

$$\widehat{f^{NW}}(x) = \frac{\sum_{i=1}^n h^{-1} Y_i K(\frac{x_i - x}{h})}{h^{-1} \sum_{i=1}^n \frac{1}{n} K(\frac{x_i - x}{h})} = \frac{\sum_{i=1}^n Y_i \frac{1}{2h} \mathbf{1}(|\frac{x_i - x}{h}| \leq 1)}{\sum_{i=1}^n \frac{1}{2h} \mathbf{1}(|\frac{x_i - x}{h}| \leq 1)} = \frac{\sum_{i:\, |x_i - x| \leq h} Y_i}{\sum_{i:\, |x_i - x| \leq h} 1}.$$

The Nadaraya-Watson estimator is an example of a linear estimator.

Definition 10.9. *Estimator $\widehat{f}(x)$ is called linear if it can be written as a linear function of y, i.e. $\widehat{f}(x) = \sum_{i=1}^n W_i(x) Y_i = W^T(x) Y$ where $Y = (y_1, \ldots, y_n)^T$, $W(x) = (w_1(x), \ldots, w_n(x))^T$ and $W(x)$ does not depend on y, only on $(x_1, \ldots, x_n)$.*

If an estimator is linear, then it is easy to find its distribution, and hence to construct a confidence interval and a confidence band (see Section 10.2.8).

Now we study the bias and the variance of the Nadaraya-Watson estimator in two frameworks, asymptotic as the sample size n grows to infinity, and for a fixed sample size.

10.2.3 Asymptotic properties of the Nadaraya-Watson estimator

As we saw in Section 10.1.4, to study consistency of an estimator, it is sufficient to study the asymptotic behaviour of its bias and variance. Thus, to study consistency of the NW estimator, we investigate asymptotic expressions for its bias and variance under the following assumptions.

Assumptions

1. Asymptotic: $n \to \infty, h \to 0$, $nh \to \infty$,
2. Design $x_1, \ldots, x_n$ is regular deterministic,
3. $x \in (0,1)$,
4. $\exists\, f''$,
5. Kernel:

$$\int_{-\infty}^{+\infty} xK(x)dx = 0, \quad 0 < \mu_2(K) \overset{def}{=} \int_{-\infty}^{+\infty} x^2 K(x)dx < \infty,$$

$$||K||_2^2 = \int_{-\infty}^{+\infty} [K(x)]^2 dx < \infty.$$

In particular, we assume that the unknown function f has a bounded second derivative and the kernel is of order 2.

A **key tool** to deriving the asymptotic expressions for the bias and the variance is approximation of a sum by an integral. Since the design (x_i) is regular deterministic, i.e. $x_i - x_{i-1} = 1/n$, for any function $g(x)$,

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) \approx \int_0^1 g(z)dz.$$

In particular, the denominator of the NW estimator is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x) \approx \int_0^1 K_h(z-x)dz &= \int_0^1 K\left(\frac{z-x}{h}\right) d\left(\frac{z}{h}\right) = \int_{\frac{0-x}{h}\rightarrow}^{\frac{1-x}{h}} K(v)dv \\ &\approx \int_{-\infty}^{+\infty} K(v)dv = 1\end{aligned}$$

since $n \to \infty$, $-x/h \to -\infty$ and $(1-x)/h \to +\infty$ as $h \to 0$. Here it is important that $x \neq 0$ and $x \neq 1$, that is, it is not at the boundary.

Asymptotic bias of the NW estimator: $b(x) \approx \frac{\mu_2(K)h^2}{2} f''(x)$.

$$\begin{aligned} b(x) = \mathbb{E}\widehat{f}(x) - f(x) &= \sum_{i=1}^{n} w_i(x)[f(X_i) - f(x)] \quad \text{[Taylor Expansion]} \\ &\approx \sum_{i=1}^{n} w_i(x)\left[f(x) + f'(x)(X_i - x) + f''(x)\frac{(X_i-x)^2}{2} - f(x)\right] \\ &= \sum_{i=1}^{n} \frac{K_h(X_i - x)}{\sum_{j=1}^{n} K_h(X_j - x)}\left[f'(x)(X_i - x) + f''(x)\frac{(X_i-x)^2}{2}\right] \\ &\approx \frac{1}{n}\left[f'(x)\sum_{i=1}^{n}(X_i - x)K_h(X_i - x) + f''(x)\sum_{i=1}^{n} K_h(X_i - x)\frac{(X_i-x)^2}{2}\right] \\ &\approx f'(x)\int_0^1 (z-x)K_h(z-x)dz + f''(x)\int_0^1 K_h(z-x)\frac{(z-x)^2}{2}dz \\ &\approx f'(x)h\int_{-x/h}^{(1-x)/h} K(v)vdv + f''(x)\frac{h^2}{2}\int_{-x/h}^{(1-x)/h} K(v)v^2dv \\ &\approx f'(x)h\int_{-\infty}^{\infty} K(v)vdv + f''(x)\frac{h^2}{2}\int_{-\infty}^{\infty} K(v)v^2dv \\ &= \frac{\mu_2(K)h^2}{2} f''(x). \end{aligned}$$

Asymptotic variance of the NW estimator: $v(x) \approx \frac{\sigma^2}{nh}||K||_2^2$:

$$\begin{aligned} v(x) &= \sigma^2 \sum_{i=1}^n [w_i(x)]^2 = \sigma^2 \sum_{i=1}^n \frac{[K_h(X_i - x)]^2}{\left[\sum_{j=1}^n K_h(X_j - x)\right]^2} \\ &\approx_{\left\{\frac{1}{n}\sum_{i=1}^n K_h(X_i - x) \approx 1\right\}} \frac{\sigma^2}{n^2} \sum_{i=1}^n [K_h(X_i - x)]^2 \\ \left\{\tfrac{1}{n}\textstyle\sum_{i=1}^n \to \int_0^1\right\} &\approx \frac{\sigma^2}{n} \int_0^1 [K_h(z-x)]^2 dz = \frac{\sigma^2}{nh} \int_0^1 \left[K\left(\frac{z-x}{h}\right)\right]^2 d\left(\frac{z-x}{h}\right) \\ \left\{v = \tfrac{z-x}{h}\right\} &= \frac{\sigma^2}{nh} \int_{-x/h}^{(1-x)/h} [K(v)]^2 dv \approx \frac{\sigma^2}{nh} \int_{-\infty}^{\infty} [K(v)]^2 dv \\ &= \frac{\sigma^2}{nh}||K||_2^2. \end{aligned}$$

Therefore, **the asymptotic MISE (AMISE)** is:

$$\begin{aligned} \text{AMISE} &= \int_0^1 \left[|b(x)|^2 + v(x)\right] dx \approx \int_0^1 \left[\frac{\mu_2(K)h^2}{2} f''(x)\right]^2 dx + \int_0^1 \frac{\sigma^2}{nh}||K||_2^2 dx \\ &= \frac{||f''||_2^2}{4} h^4 [\mu_2(K)]^2 + \frac{\sigma^2}{n} \frac{||K||_2^2}{h}. \end{aligned}$$

We are in general interested in having the "best" estimator of the function. This can be interpreted as finding h and K that minimise this error. We start with optimising over the kernel, introducing canonical kernels.

10.2.4 Canonical Kernel

Given a kernel $K(x)$ of order 2, consider a scale family of kernels:

$$\left\{K_\delta(x) = \frac{1}{\delta} K\left(\frac{x}{\delta}\right), \delta > 0\right\}$$

Definition 10.10. *The* **canonical bandwidth**, *δ_0, is defined by*

$$\delta_0 = \left(\frac{||K||_2^2}{[\mu_2(K)]^2}\right)^{\frac{1}{5}},$$

where $\mu_2(K) = \int_{-\infty}^{+\infty} x^2 K(x) dx$ and $||K||_2 = \sqrt{\int_{-\infty}^{+\infty} [K(x)]^2 dx}$.

Then, given a scale family of kernels $\left\{K_\delta(x) = \frac{1}{\delta} K\left(\frac{x}{\delta}\right), \delta > 0\right\}$, the **canonical kernel**, K_{δ_0}, is

$$K_{\delta_0}(x) = \frac{1}{\delta_0} K\left(\frac{x}{\delta_0}\right).$$

Choosing the canonical kernel in the scale family allows comparison across families of kernels. For example, we shall see that if we choose a canonical kernel, the optimal bandwidth does not depend on the kernel.

Lemma 10.1. *For a scale family $\{K_\delta, \delta > 0\}$, the canonical bandwidth δ_0 satisfies*

$$||K_{\delta_0}||_2^2 = [\mu_2(K_{\delta_0})]^2.$$

Proof. We show that if $||K_h||_2^2 = [\mu_2(K_h)]^2$ if and only if $h = \delta_0$. Consider separately the right and left hand sides.

$$||K_h||_2^2 = \int_{-\infty}^{\infty} [K_h(x)]^2 \, dx = \frac{1}{h}\int_{-\infty}^{\infty} \left[K\left(\frac{x}{h}\right)\right]^2 d\left(\frac{x}{h}\right) = \frac{1}{h}||K||_2^2$$

$$\mu_2(K_h) = \int_{-\infty}^{+\infty} x^2 K_h(x) dx = h^2 \int_{-\infty}^{+\infty} \left(\frac{x}{h}\right)^2 K\left(\frac{x}{h}\right) d\frac{x}{h} = h^2\mu_2(K)$$

Therefore, $||K_h||_2^2 = \mu_2(K_h)^2 \Leftrightarrow \frac{1}{h}||K||_2^2 = [h^2\mu_2(K)]^2$ which implies that

$$h = \left(\frac{||K||_2^2}{[\mu_2(K)]^2}\right)^{\frac{1}{5}} = \delta_0.$$

□

10.2.5 Optimal kernel and optimal bandwidth

We are looking for the kernel and the bandwidth that minimise the asymptotic MISE. The AMISE is given by

$$\text{AMISE} \approx \frac{||f''(x)||_2^2}{4}\left[h^2\mu_2(K)\right]^2 + \frac{\sigma^2}{n}\frac{||K||_2^2}{h}.$$

For a canonical kernel, the AMISE factorises into a term that depends on bandwidth and a term that depends on the kernel:

$$\text{AMISE} \approx ||K||_2^2\left[h^4\frac{||f''(x)||_2^2}{4} + h^{-1}\frac{\sigma^2}{n}\right].$$

For any kernel, we can also define the **optimal bandwidth**, h_{opt}, by minimising the AMISE over h. First, we take a derivative of the AMISE with respect to h:

$$\frac{\partial}{\partial h}\text{AMISE} = \left[4h^3 C_1 - h^{-2}\frac{C_2}{n}\right] = 0$$

where $C_1 = ||f''(x)||_2^2\mu_2(K)^2/4$, and $C_2 = \sigma^2||K||_2^2$, which is solved by

$$h_{\text{opt}} = \left(\frac{C_2}{4nC_1}\right)^{\frac{1}{5}} = \left(\frac{\sigma^2||K||_2^2}{n||f''(x)||_2^2\mu_2(K)^2}\right)^{\frac{1}{5}}$$

which corresponds to the minimum of AMISE. For a canonical kernel we note that $||K||_2^2 = \mu_2(K)^2$ and so the optimal bandwidth does not depend on the kernel but it does depend on the unknown function.

Using the optimal bandwidth, the AMISE becomes

$$\text{AMISE} = \frac{5\sigma^{\frac{8}{5}}||f''(x)||_2^{\frac{2}{5}}}{4n^{\frac{4}{5}}}\left(\sqrt{\mu_2(K)}||K||_2^2\right)^{\frac{4}{5}}.$$

Optimal kernel: choose the kernel K to minimize the AMISE. From the preceding expression, this corresponds to minimising the quantity $\sqrt{\mu_2(K)}||K||_2^2$. We note that this is independent of bandwidth, in the sense that $\sqrt{\mu_2(K)}||K||_2^2 = \sqrt{\mu_2(K_\delta)}||K\delta||_2^2$ for all δ. However, rescaling by δ in this way will change the corresponding optimal bandwidth, so that the rescaled kernel with its optimal bandwidth is unchanged. We can use this freedom to set $\mu_2(K) = 1$ (which requires rescaling by $\delta = 1/\sqrt{\mu_2(K)}$). For this choice, minimising the bandwidth-optimised AMISE is equivalent to minimising $||K||_2^2$ under the constraints:

$$\int K(x)dx = 1, \quad \int xK(x)dx = 0, \quad \int x^2 K(x)dx = 1.$$

The canonical kernel that minimises $||K||_2$ under these constraints is

$$K^{\text{opt}}(x) = \frac{3}{4}\frac{1}{\sqrt{5}}\left(1 - \frac{x^2}{5}\right)\mathbf{1}(|x| \leq \sqrt{5}).$$

This kernel is called the Epanechnikov kernel. For the Epanechnikov kernel, $||K||_2^2 = 3/5\sqrt{5}$ and $\mu_2(K) = 1$ by construction, so the optimal bandwidth is

$$h_{\text{opt}} = \left(\frac{3\sigma^2}{5\sqrt{5}n||f''(x)||_2^2}\right)^{\frac{1}{5}}.$$

Therefore, the **optimal kernel with the optimal bandwidth**, $K_{h_{\text{opt}}}$, is given by

$$K_{h_{\text{opt}}}(x) = \frac{1}{h_{\text{opt}}}K\left(\frac{x}{h_{\text{opt}}}\right) = \frac{3}{4}\frac{1}{\sqrt{5}h_{\text{opt}}}\left(1 - \frac{x^2}{5h_{\text{opt}}^2}\right)\mathbf{1}(|x| \leq \sqrt{5}h_{\text{opt}}),$$

and the Nadaraya-Watson estimator constructed with this kernel has the smallest AMISE.

The efficiency of a kernel family $\{K_\delta,\, \delta > 0\}$ for a given kernel K is defined as

$$\frac{\sqrt{\mu_2(K)}\,||K||_2^2}{\sqrt{\mu_2(K^{\text{opt}})}||K^{\text{opt}}||_2^2} = \frac{\sqrt{\mu_2(K_{\delta_0})}||K_{\delta_0}||_2^2}{\sqrt{\mu_2(K^{\text{opt}}_{\delta_0^{\text{opt}}})}||K^{\text{opt}}_{\delta_0^{\text{opt}}}||_2^2} = \left(\frac{\mu_2(K_{\delta_0})}{\mu_2(K^{\text{opt}}_{\delta_0^{\text{opt}}})}\right)^{\frac{5}{2}} = \left(\frac{||K_{\delta_0}||_2^2}{||K^{\text{opt}}_{\delta_0^{\text{opt}}}||_2^2}\right)^{\frac{5}{4}}$$

where δ_0 is the canonical bandwidth for this kernel family, K^{opt} is the Epanechnikov kernel and δ_0^{opt} is its canonical bandwidth. The efficiency to the fourth fifths power gives the ratio of the AMISE for this family of kernels relative to the optimal kernel family. For many kernel families, the efficiencies are close to 1, for instance, it is 0.951 for the Gaussian kernel family, 0.930 for the box kernel family and 0.986 for the triangular kernel family.

Note that since the optimal bandwidth depends on the unknown function, this expression gives a theoretical bound but it is not applicable in practice. One way to avoid dependency on the unknown function is to take $h_{\text{opt}} = Cn^{-1/5}$ which gives the same order of MISE in n but not the optimal constant. Another way to find the best h that is used in practice is to use another approximation of MISE which results in the approach called cross-validation.

10.2.6 Non-asymptotic properties of the Nadaraya-Watson estimator

Nonasymptotic properties of the Nadaraya-Watson estimator can be found in the form of upper bounds on the absolute value of the bias and the variance, and hence on the MSE and MISE. We shall see that the upper bounds are the same functions of the sample size n. The constants in the upper bounds inform us how the errors depend on other features of the model, such as the kernel, the smoothness of the function, design, etc.

Before we state the upper bounds, we will define a class of smooth functions, the Hölder Class $\boldsymbol{H^\beta(M)}$. When the parameter β is an integer, the class $\boldsymbol{H^\beta(M)}$ contains functions with β derivatives whose absolute values are bounded by M. However, the class is defined for arbitrary values $\beta > 0$.

Definition 10.11. *The Hölder Class* $\boldsymbol{H^\beta(M)}$ *of functions on* $[0,1]$ *with* $\beta > 0,\ M > 0$ *is defined as the set of functions* f *that satisfy the following conditions with* $k = \lfloor \beta \rfloor$*:*

1. $|f^{(k)}(x)| \leq M$ *for all* $x \in [0,1]$,
2. $|f^{(k)}(x) - f^{(k)}(y)| \leqslant M|x-y|^{\beta-k},\ \forall x, y \in [0,1]$,
   where $f^{(k)}$ *is the* k*th derivative of* f.

If $\beta \in (0,1)$, $k = 0$ *and* $f^{(0)}(x) = f(x)$.

Example: if $\beta = 1$, the Hölder class $\boldsymbol{H^1(M)}$ contains functions such that $|f'(x)| \leq M$ for all $x \in [0,1]$.

Example: the function $f(x) = \sqrt{|x-0.5|}$, $x \in [0,1]$, does not have a derivative for all $x \in [0,1]$ but it belongs to the Hölder class $\boldsymbol{H^\beta(M)}$ with $\beta = 1/2$ and $M = 1$ due to the inequality

$$|\sqrt{|z|} - \sqrt{|y|}| \leqslant \sqrt{|z-y|} \quad \forall z, y \in [0,1].$$

Now we derive upper bounds on the absolute value of the bias and the variance of the Nadaraya-Watson estimator of a function f that belongs to a Hölder class $\boldsymbol{H^\beta(M)}$ with $\beta \in (0,1)$.

Proposition 10.2. *Suppose that* $f \in \mathbb{H}^\beta(M)$ *on* $[0,1]$*, with* $\beta \in (0,1]$ *and* $M > 0$*. Let* $\widehat{f}_n^{NW}$ *be the Nadaraya – Watson estimator of* f*.*

Assume also that:

a) *the design* $(X_1, \ldots, X_n)$ *is regular deterministic;*

b) $var(\varepsilon_i) = \sigma^2$*;*

c) $\exists\, \lambda_0 > 0$ *such that* $\forall\, x \in [0,1]$,

$$\frac{1}{n}\sum_{i=1}^{n} K_h\left(X_i - x\right) \geq \lambda_0;$$

d) $supp(K) \subseteq [-1,1]$ *(i.e.* $K(x) = 0$ *for* $x \notin [-1,1]$*),*
   and $\exists\, K_{\max} \in (0,\infty)$ *such that* $0 \leq K(u) \leq K_{\max}$, $\forall u \in \mathbb{R}$.

Then, for all $x_0 \in [0,1]$ *and* $h \geq 1/(2n)$,

$$|b(x_0)| \leq Mh^\beta, \quad v(x_0) \leq \frac{\sigma^2 K_{\max}}{nh\lambda_0}.$$

Proof. 1. The bias of the NW estimator when $f \in H^B(M)$ with $\beta \in (0,1)$ is:

$$\text{bias}(\widehat{f^{NW}}(x)) = \mathbb{E}(\widehat{f^{NW}}(x)) - f(x) = \sum_{i=1}^{n} W_i^{NW}(x)\,[f(x_i) - f(x)]\,.$$

Note that $\forall\, x$, $\sum_{i=1}^n W_i^{NW}(x) = 1$, since

$$\sum W_i(x) = \frac{\sum_{i=1}^n K_h(x_i - x)}{\sum_{i=1}^n K_h(x_i - x)} \mathbf{1}\left(\sum K_h(x_i - x) \neq 0\right) = 1.$$

Therefore, the bias is given by

$$\text{bias}(\widehat{f}^{NW}(x)) = \sum_{i=1}^{n} W_i^{NW}(x)[f(x_i) - f(x)].$$

Since the support of K is $[-1,1]$, the support of $K_h(x) = \frac{1}{h}K(x/h)$ is $[-h,h]$, therefore the sum is only over those i where $|x_i - x| \leq h$, that is,

$$\begin{aligned}
|\text{bias}(\widehat{f}^{NW}(x))| &= \frac{\left|\sum_i K(\frac{x_i - x}{h})(f(x_i) - f(x))\right|}{\sum_i K(\frac{x_i-x}{h})} = \frac{\left|\sum_{i:\,|x_i - x| \leq h} K(\frac{x_i-x}{h})[f(x_i) - f(x)]\right|}{\sum_i K(\frac{x_i-x}{h})} \\
&\leq \frac{\sum_{i\,:|x_i - x| \leq h} K(\frac{x_i-x}{h})\,|f(x_i) - f(x)|}{\sum_i K(\frac{x_i-x}{h})} \leq \frac{\sum_{i\,:|x_i-x|\leq h} K(\frac{x_i-x}{h})\, M|x_i - x|^\beta}{\sum_i K(\frac{x_i-x}{h})} \\
&\leq M h^\beta,
\end{aligned}$$

using $K(z) \geq 0$ for all z. In particular, the bias is small when h is small, that is, $\text{bias}(\widehat{f}^{NW}(x)) \to 0$ if $h \to 0$. The extension of the proof to $\beta = 1$ is left as an exercise.

2. The variance of the NW estimator can be written as

$$v(x) = \text{Var}(\widehat{f}_n^{NW}(x)) = \text{Var}\left(\sum_{i=1}^n w_i(x)(Y_i)\right) = \sum_{i=1}^n [w_i(x)]^2 \text{Var}(Y_i)$$

since the Y_i's are independent. From assumptions (a) & (b), we know that $\text{Var}(Y_i) = \sigma^2$, since the x_i's are fixed. Therefore,

$$\begin{aligned}
v(x) &= \sigma^2 \sum_{i=1}^n \frac{[K_h(X_i - x)]^2}{\left[\sum_{j=1}^n K_h(X_j - x)\right]^2} && \text{assumption d) } K(z) \geq 0 \text{ for all } z \\
&\leqslant \sigma^2 \frac{\frac{K_{\max}}{h}\sum_{i=1}^n K_h(X_i - x)}{\left[\sum_{j=1}^n K_h(X_j - x)\right]^2} && \text{assumption d), } \forall u, K(u) \leqslant K_{\max} \text{ implies } K_h(X_i - x) = \tfrac{1}{h}K\left(\tfrac{X_i - x}{h}\right) \leqslant \tfrac{K_{\max}}{h} \\
&\leqslant \sigma^2 \frac{\frac{K_{\max}}{h}\sum_{i=1}^n K_h(X_i - x)}{n\lambda_0 \sum_{j=1}^n K_h(X_j - x)} && \text{assumption c) } \exists \lambda_0 > 0 \text{ such that } \forall x \in [0,1], \textstyle\sum_{i=1}^n K_h(X_i - x) \geqslant n\lambda_0. \\
&= \frac{\sigma^2 K_{\max}}{nh\lambda_0}
\end{aligned}$$

□

Now we consider the bounds on the MSE of the NW estimator. Under the conditions of Proposition 10.2,

$$\mathrm{MSE}(\widehat{f}_n^{NW}(x_0)) = [\mathrm{bias}(\widehat{f}_n^{NW}(x_0))]^2 + \mathrm{Var}(\widehat{f}_n^{NW}(x_0)) \leq M^2 h^{2\beta} + \frac{\sigma^2 K_{\max}}{nh\lambda_0}.$$

The upper bound on MSE is the smallest if

$$h = h_n = \left(\frac{\sigma^2 K_{\max}}{2\beta M^2 \lambda_0 n}\right)^{1/(2\beta+1)},$$

and the corresponding MSE bound is

$$\begin{aligned}\mathrm{MSE}(\widehat{f}_{n,h_n}^{NW}(x_0)) &\leq M^2 \left(\frac{\sigma^2 K_{\max}}{2\beta M^2 \lambda_0 n}\right)^{2\beta/(2\beta+1)} + \frac{\sigma^2 K_{\max}}{n\lambda_0}\left(\frac{2\beta M^2 \lambda_0 n}{\sigma^2 K_{\max}}\right)^{1/(2\beta+1)} \\ &\leq (1+2\beta) M^{2/(2\beta+1)} \left(\frac{\sigma^2 K_{\max}}{2\beta\lambda_0 n}\right)^{2\beta/(2\beta+1)} \to 0 \quad \text{as} \quad n \to \infty.\end{aligned}$$

Hence, the Nadaraya – Watson estimator with $h = h_n = \left(\frac{\sigma^2 K_{\max}}{2\beta M^2 \lambda_0 n}\right)^{1/(2\beta+1)}$ and kernel K satisfying conditions of Proposition 10.2, is consistent for estimating functions from Hölder class $\mathbb{H}^{\beta}(M)$ for $\beta \in (0, 1]$.

Example 10.6. *(continued) Derive upper bounds on the absolute value of the bias and the variance of the NW estimator with the box kernel $K(z) = \frac{1}{2}\mathbf{1}(z \in [-1, 1])$ under the nonparametric regression model with $\sigma^2 = 1$ and $x_i = i/n$. Let $f \in H^{\beta}(M)$, $M = 5$,$\beta = 1/2$.*

Now we verify the assumptions of Proposition 10.2. Assumptions $a), b)$ are satisfied. Assumption c) is $\frac{1}{2n}\sum_{i:|x_i - x|\leq h} \frac{1}{h} \geq \lambda_0$, $h \geq 1/2n$.

Let's count the number of integers i between 1 and n such that $|i/n - x| \leq h$. Since

$$|i/n - x| \leq h \Leftrightarrow (nx - nh) \leq i \leq (nx + nh),$$

we need to count the number of integers in the interval $[nx - nh, nx + nh]$.

In general, in an interval $[a, a + b]$ for some $b > 0$, the number of of integers is $\lfloor b \rfloor$ if a is not integer, and it is $\lfloor b \rfloor$ +1 if a is integer. Here $\lfloor b \rfloor$ is the lower integer part of b, that is, the largest integer that is less than or equal to b, e.g. $\lfloor 5 \rfloor = 5$, $\lfloor 7.3 \rfloor = 7$ and $\lfloor 2.8 \rfloor = 2$.

Therefore, the smallest number of integers in the interval $[nx - nh, nx + nh]$ is $\lfloor 2nh \rfloor$ which is greater than $2nh - 1$ since $\lfloor 2nh \rfloor \leq 2nh < \lfloor 2nh \rfloor + 1$ by the definition of the lower integer part. Hence, we need $h > 1/(2n)$, and then we can take $\lambda_0 = 1 - 1/(2nh) > 0$ since

$$\frac{1}{2n}\sum_{i:|x_i - x|\leq h} \frac{1}{h} \geq \frac{2nh - 1}{2nh} = 1 - 1/(2nh) = \lambda_0$$

Assumption d) is satisfied with $K_{\max} = 1/2$.

Therefore, by Proposition 10.2, for $n = 12$ and $h > 1/24$,

$$|b(x)| \leq Mh^{\beta} = 5\sqrt{h}, \quad v(x) \leq \frac{1}{2nh(1 - 1/(2nh))} = \frac{1}{2nh - 1} = \frac{1}{24h - 1}.$$

The corresponding MSE (and MISE) for $\hat{f}^{NW}(x)$ is bounded by

$$MSE(\hat{f}^{NW}(x)) = b^2(x) + v(x) \leq 25h + \frac{1}{24h - 1}.$$

The derivative of the upper bound with respect to h is

$$25 - \frac{24}{(24h-1)^2}$$

which is zero for $h > 1/24$ at

$$h_{opt} = \frac{1}{24}\left(1 + \sqrt{\frac{24}{25}}\right) = 0.0825.$$

This corresponds to the minimum of the MSE since the second derivative with respect to h of the upper bound is $\frac{2 \cdot 24^2}{(24h-1)^3}$ which is positive.

Therefore, the optimal bandwidth is 0.0825*.*

10.2.7 Rates of convergence

We would like to find the estimator of f which is not only consistent, but also achieves the best possible rate of convergence over some class of functions $\mathcal{F}$, such as the Hölder class $H^\beta(M)$. Now we determine the rate of convergence of the NW estimator, in both local and global distances, and address the question whether it is possible to achieve a faster rate of convergence.

Definition 10.12. *ϕ_n is the* **convergence rate of an estimator $\widehat{f}_n$ at point** x_0 *(local rate of convergence) over a class of functions $\mathcal{F}$, if*

$$0 < c \leqslant \sup_{f \in \mathcal{F}} \mathbb{E}\left[\frac{|\widehat{f}_n(x_0) - f(x_0)|}{\phi_n}\right]^2 \leqslant C < \infty,$$

where constants c and C do not dependent on n, and the rate ϕ_n is only related to n and the function class $\mathcal{F}$.

Similarly, the global rate of convergence of estimator $\widehat{f}_n$ over a class of functions $\mathcal{F}$ is ϕ_n if

$$0 < c \leqslant \sup_{f \in \mathcal{F}} \mathbb{E}\left[\frac{||\widehat{f}_n - f||_2}{\phi_n}\right]^2 \leqslant C < \infty,$$

where the constants c and C do not depend on n, and the rate ϕ_n is only related to n and the function class $\mathcal{F}$.

Recall that $||\widehat{f}_n(x) - f(x)||_2 = \sqrt{\int_0^1 [\widehat{f}_n(x) - f(x)]^2 dx}$.

Definition 10.13. *For a class of functions $\mathcal{F}$, $\phi_n^\star$ is the* **local minimax convergence rate**, *if*

$$0 < c \leqslant \inf_{\widehat{f}_n} \sup_{x_0 \in (0,1)} \sup_{f \in \mathcal{F}} \mathbb{E}\left[\frac{|\widehat{f}_n(x_0) - f(x_0)|}{\phi_n^\star}\right]^2 = \inf_{\widehat{f}_n} \sup_{x_0 \in (0,1)} \sup_{f \in \mathcal{F}} \frac{MSE(\widehat{f}_n(x_0))}{(\phi_n^\star)^2} \leqslant C < \infty,$$

where the constants c and C do not depend on n, and the rate $\phi_n^\star$ is only related to n and the function class $\mathcal{F}$.

Similarly, for a class of functions $\mathcal{F}$, $\phi_n^\star$ is the **global minimax convergence rate***, if*

$$0 < c \leqslant \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}\left[\frac{||\widehat{f}_n - f||_2}{\phi_n^\star}\right]^2 = \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \frac{MISE(\widehat{f}_n)}{(\phi_n^\star)^2} \leqslant C < \infty,$$

where constants c and C do not depend on n, and the rate $\phi_n^\star$ is only related to n and the function class $\mathcal{F}$.

Definition 10.14. *An estimator $\widehat{f}_n$ is said to achieve a minimax rate of convergence (local or global), if the rate of convergence of this estimator is the corresponding (local or global) minimax rate of convergence.*

Now we investigate whether the local rate of convergence for the Nadaraya-Watson estimator is minimax.

Theorem 10.1. *Let assumptions of Proposition 10.2 hold for all $x \in [0, 1]$. Then, the NW estimator $\widehat{f}^{NW}(x)$ with $h = \alpha n^{-1/(2\beta+1)}$ for same $\alpha > 0$ satisfies*

$$\lim_{n\to\infty} \sup_{x_0 \in [0,1]} \sup_{f \in H^\beta(M)} \mathbb{E}\left[\left((\widehat{f}_n^{NW}(x_0) - f(x_0))n^{\beta/(2\beta+1)}\right)^2\right] \leq C < \infty,$$

where constant C depends only on $\beta, M, \sigma^2, \lambda_0, K_{\max}, \alpha$.

Proof. By Proposition 10.2, $\forall f \in H^\beta(M), \forall x \in [0, 1]$,

$$\mathbb{E}\left[\left(\widehat{f}_n^{NW}(x) - f(x)\right)^2\right] \leq C n^{\frac{-2\beta}{2\beta+1}}$$

with $C < \infty$ dependent on $K_{\max}, \lambda_0, \beta, M, \alpha, \sigma^2$ which can be written as

$$\mathbb{E}\left[\left((\widehat{f}_n^{NW}(x) - f(x))n^{\beta/2\beta+1}\right)^2\right] \leq C.$$

Taking supremum over $f \in H^\beta(M)$, $x \in [0, 1]$ and n, as $n \to \infty$, we have the statement. □

Therefore, the pointwise rate of convergence of the Nadaraya-Watson estimator is $n^{-\beta/(2\beta+1)}$. In fact, it can be shown (Tsybakov, 2009, chapter 2) that this is the local minimax rate of convergence, so the Nadaraya-Watson estimator achieves this minimax rate and so it is in this sense the “best” estimator, but there do exist other estimators that achieve this rate of convergence. It is straightforward to show that the NW estimator also achieves the global minimax rate of convergence.

The upper bounds being used here apply for the Hölder space with $\beta \in (0, 1]$. For the Nadaraya-Watson estimator to achieve the minimax convergence rate for $\beta > 1$, one needs to use kernels of higher order. **Local polynomial estimators**, which will be discussed in Section 10.2.12 are locally and globally minimax for $\beta > 1$.

10.2.8 Inference using a linear estimator

In this subsection we consider the nonparametric regression model

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

with independent errors $\varepsilon_i \sim N(0, \sigma^2)$ and a deterministic design $(x_1, \ldots, x_n)$. These assumptions imply that $\mathbb{E}(Y_i) = f(X_i)$ and $\mathrm{Var}(Y_i) = \sigma^2$.

10.2.9 Confidence intervals for $f(x_0)$ based on a linear estimator

Denote $b(x) = \text{bias}(\widehat{f}(x)) = \mathbb{E}\left[\widehat{f}(x) - f(x)\right]$ and $v(x) = \text{Var}(\widehat{f}(x))$. Then, for a **linear estimator** $\widehat{f}(x) = \sum_{i=1}^{n} Y_i w_i(x)$,

$$\begin{aligned}\mathbb{E}\left(\widehat{f}(x)\right) &= \sum_{i=1}^{n} f(x_i)w_i(x) = b(x) + f(x)\\ \text{Var}\left(\widehat{f}(x)\right) &= \sigma^2 \sum_{i=1}^{n} [w_i(x)]^2 = v(x),\end{aligned}$$

therefore $\widehat{f}(x) \sim N\left(b(x) + f(x), v(x)\right)$.

The variance depends on the weights $w_i(x)$ and σ which are known, so it can be calculated exactly. If we knew the bias, which depends on the unknown function, we could construct $(1-\alpha)100\%$ confidence interval using the fact that the following inequality

$$-z_{\frac{\alpha}{2}} \leqslant \frac{\widehat{f}(x) - [b(x) + f(x)]}{\sqrt{v(x)}} \leqslant z_{\frac{\alpha}{2}}$$

holds with probability $1-\alpha$, that is,

$$f(x) \in [\widehat{f}(x) - b(x) - z_{\frac{\alpha}{2}}\sqrt{v(x)}, \widehat{f}(x) - b(x) + z_{\frac{\alpha}{2}}\sqrt{v(x)}].$$

Here $z_\alpha = \Phi^{-1}(1-\alpha)$ where $\Phi(x)$ is the cumulative distribution function of $N(0,1)$.

However, the bias is unknown, so it is not possible to construct the exact confidence interval. There are two approaches to addressing this issue. The first one is to construct an asymptotic confidence interval where the estimator is constructed in such a way that asymptotically the bias is much smaller than the variance, and therefore may be treated as 0. For the NW estimator, this means choosing a smaller bandwidth. The second one is to use an upper bound on the bias to construct a conservative confidence interval.

- $(1-\alpha)100\%$ Conservative Confidence Interval for $f(x)$.

  If $|b(x)| \leqslant b_0(x)$ & $v(x) \leqslant v_0(x)$, then

  $$f(x) \in \widehat{f}(x) \pm \left(b_0(x) + z_{\frac{\alpha}{2}}\sqrt{v_0(x)}\right).$$

- $(1-\alpha)100\%$ Asymptotic Confidence Interval for $f(x)$.

  Choose the estimator $\widehat{f}(x)$ so that $b(x)^2 \ll v(x)$, thus we can assume $b(x) \approx 0$:

  $$f(x) \in \widehat{f}(x) \pm z_{\frac{\alpha}{2}}\sqrt{v(x)}.$$

  The asymptotic expression for the variance is often used in this case.

10.2.10 Confidence intervals using the Nadaraya-Watson estimator

For a Nadaraya-Watson estimator $f \in H^\beta(M)$ on $x \in [0,1]$, under the conditions of Proposition 10.2,

$$v(x) \leqslant \frac{\sigma^2 K_{\max}}{nh\lambda_0}, \quad |b(x)| \leqslant Mh^\beta.$$

Therefore, a $(1-\alpha)100\%$ **Conservative Confidence Interval** for $f(x)$ is

$$\widehat{f}^{NW}(x) \pm \left(Mh^\beta + z_{\alpha/2}\sigma\sqrt{K_{\max}/(nh\lambda_0)}\right)$$
$$= \left[\widehat{f}^{NW}(x) - Mh^\beta - z_{\alpha/2}\sigma\sqrt{K_{\max}/(nh\lambda_0)}, \widehat{f}^{NW}(x) + Mh^\beta + z_{\alpha/2}\sigma\sqrt{K_{\max}/(nh\lambda_0)}\right].$$

Alternatively, taking the limit $n \to \infty$ and $h \to 0$,

$$v(x) \approx \frac{\sigma^2}{nh}||K||_2^2, \quad b(x) \approx \frac{\mu_2(K)h^2}{2} f''(x) \approx 0.$$

Therefore, a $(1-\alpha)100\%$ **Asymptotic Confidence Interval** for $f(x)$ is

$$\widehat{f}^{NW}(x) \pm z_{\alpha/2}\sigma\sqrt{||K||_2^2/(nh))}$$
$$= \left[\widehat{f}^{NW}(x) - z_{\alpha/2}\sigma\sqrt{||K||_2^2/(nh)}, \widehat{f}^{NW}(x) + z_{\alpha/2}\sigma\sqrt{||K||_2^2/(nh)}\right].$$

10.2.11 Asymptotic Confidence Band for f

Assume that the bias of $\widehat{f}(x)$ is much smaller than its standard deviation and is close to 0, i.e. $|b(x)| \ll \sqrt{v(x)}$ and $b(x) \approx 0$. Then, an asymptotic $(1-\alpha)100\%$ confidence band based on the NW estimator is given by

$$\left\{f : |f(x) - \widehat{f}(x)| \leqslant c_\alpha\sqrt{v(x)},\ \forall\, x \in [a,b]\right\}$$

with

$$c_\alpha \approx \sqrt{2\log\left(\frac{a_0}{\alpha h}\right)}, \text{ where } a_0 = \frac{|b-a|}{\pi}\frac{||K'||_2}{||K||_2},$$

(see Wasserman, section 5.7). For the NW estimator, we can use $v(x) \approx \frac{\sigma^2}{nh}||K||_2^2$.

Confidence bands can be used to test hypotheses about f, e.g.

$$H_0 : f(x) = \text{constant}\, \forall x \in [0,1].$$

10.2.12 Local polynomial estimators.

Motivation and definition The Nadaraya-Watson estimator can be viewed as a local constant least squares approximation of the unknown function. If the kernel K takes only nonnegative values, then for each $x \in [0,1]$, $\widehat{f}_n^{NW}(x)$ satisfies

$$\widehat{f}_n^{NW}(x) = \arg\min_{\theta_x \in \mathbb{R}}\left\{\sum_{i=1}^n (Y_i - \theta_x)^2 K\left(\frac{X_i - x}{h}\right)\right\}$$
$$= \arg\min_{\theta_x \in \mathbb{R}}\left\{\sum_{i=1}^n (\theta_x^2 - 2\theta_x Y_i + Y_i^2) K\left(\frac{X_i - x}{h}\right)\right\}$$
$$= \arg\min_{\theta_x \in \mathbb{R}}\left\{\theta_x^2 \cdot \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) - \theta_x \cdot 2\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right) + C_{X_i,Y_i}(x)\right\}$$

Therefore, if $\sum_{j=1}^{n} K_h(X_j - x) \neq 0$, the value of θ_x that minimises this weighed sum of squares coincides with the Nadaraya-Watson estimator:

$$f_n^{NW}(x) = \frac{\sum_{i=1}^{n} Y_i K_h(X_i - x)}{\sum_{j=1}^{n} K_h(X_j - x)}.$$

This estimator can be generalised further by considering a local polynomial rather than a local constant approximation. For a function $f(x)$, if $\exists f^{(k)}(x)$, then for x_i sufficiently close to x,

$$\begin{aligned} f(x_i) \approx & f(x) + f'(x)(x_i - x) + \cdots + \frac{f^{(k)}(x)}{k!}(x_i - x)^k = \sum_{j=0}^{k} \frac{f^{(j)}(x)}{j!}(x_i - x)^j \\ = & \sum_{j=0}^{k} \left[f^{(j)}(x)h^j\right] \left[\frac{1}{j!}\left(\frac{x_i - x}{h}\right)^j\right] = U_{x,i}^T \theta_x \end{aligned}$$

where

$$\begin{aligned} \theta_x &= \left(f(x), f'(x)h, f''(x)h^2, \; \dots \; , f^{(k)}(x)h^k\right)^T \\ U_{x,i} &= \left(1, \frac{x_i - x}{h}, \frac{1}{2!}\left(\frac{x_i - x}{h}\right)^2, \dots, \frac{1}{k!}\left(\frac{x_i - x}{h}\right)^k\right)^T \end{aligned}$$

Definition 10.15. *A local polynomial estimator of $f(x)$ of order k , denoted $LP(k)$ estimator, is defined by*

$$\widehat{f}_n^{LP}(x) = \widehat{\theta}_0(x)$$

where for each x $\widehat{\theta}(x) = \left(\widehat{\theta}_0(x), \widehat{\theta}_1(x), \dots, \widehat{\theta}_k(x)\right)^T$ is the solution of

$$\widehat{\theta}(x) = \arg \min_{\theta_x \in \mathbb{R}^{k+1}} \left\{ \sum_{i=1}^{n} (Y_i - U_{x,i}^T \theta_x)^2 K\left(\frac{X_i - x}{h}\right) \right\}.$$

For each $m = 1, \dots, k$, $\widehat{\theta}_m(x)/h^m$ is an estimator of $f^{(m)}(x)$.

Therefore, the local polynomial estimator provides simultaneous estimators not only for $f(x)$ but also for all existing derivatives of f.

This estimator can be written explicitly. Noticing that the expression to be minimised is quadratic in the vector θ_x, we can open the brackets to obtain

$$\begin{aligned} \widehat{\theta}_x &= \arg\min_{\theta_x} \left\{ \sum_{i=1}^{n} (Y_i - U_{x,i}^T \theta_x)^2 K\left(\frac{X_i - x}{h}\right) \right\} \\ &= \arg\min_{\theta_x} \left\{ \sum_{i=1}^{n} (\theta_x^T U_{x,i} U_{x,i}^T \theta_x - 2U_{x,i}^T \theta_x Y_i + Y_i^2) K\left(\frac{X_i - x}{h}\right) \right\} \\ &= \arg\min_{\theta_x} \left\{ \theta_x^T \cdot \sum_{i=1}^{n} U_{x,i} U_{x,i}^T K\left(\frac{X_i - x}{h}\right) \cdot \theta_x - \theta_x^T \cdot 2 \sum_{i=1}^{n} Y_i U_{x,i} K\left(\frac{X_i - x}{h}\right) + C_{X_i,Y_i}(x) \right\} \end{aligned}$$

which is equivalent to

$$\widehat{\theta}_x = \arg\min_{\theta_x} \left\{ \theta_x^T \cdot B(x) \cdot \theta_x - 2\theta_x^T \cdot a(x) \right\}$$

where the matrix $B(x)$ and vector $a(x)$ are defined by

$$B(x) = \frac{1}{nh}\sum_{i=1}^{n} U_{x,i}U_{x,i}^T K\left(\frac{X_i - x}{h}\right) \qquad a(x) = \frac{1}{nh}\sum_{i=1}^{n} Y_i U_{x,i} K\left(\frac{X_i - x}{h}\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n} U_{x,i}U_{x,i}^T K_h(X_i - x) \qquad = \frac{1}{n}\sum_{i=1}^{n} Y_i U_{x,i} K_h\left(X_i - x\right)$$

Hence, if $B(x)$ is invertible,

$$\widehat{\theta}_x = B^{-1}(x)a(x).$$

Therefore, the Local Polynomial estimator can be written as

$$\widehat{f}_n^{LP}(x) = \widehat{\theta}_0(x) = e_1^T B^{-1}(x)a(x)$$

where the matrix $B(x)$ and vector $a(x)$ are defined above and $e_1^T = (1, 0, 0, \cdots, 0)$.

Note that the local polynomial estimator $\widehat{f}_n^{LP}(x)$ is **linear**:

$$\begin{aligned} f_n^{LP}(x) &= e_1^T B^{-1}(x)a(x) = e_1^T B^{-1}(x) \cdot \frac{1}{n}\sum_{i=1}^{n} Y_i U_{x,i} K_h\left(X_i - x\right) \\ &= \sum_{i=1}^{n} Y_i \cdot \frac{1}{n} K_h\left(X_i - x\right) \sum_{j=0}^{k} [B^{-1}(x)]_{0,j} \frac{1}{j!}\left(\frac{x_i - x}{h}\right)^j \\ &= \sum_{i=1}^{n} Y_i w_i(x) \end{aligned}$$

with weights

$$w_i(x) = \frac{1}{n} K_h\left(X_i - x\right) \sum_{j=0}^{k} [B^{-1}(x)]_{0,j} \frac{1}{j!}\left(\frac{x_i - x}{h}\right)^j$$

that are independent of $Y_1, \ldots, Y_n$.

Bias, variance, consistency and the rate of convergence for local polynomial estimator

Proposition 10.3. *Suppose that $f \in H^\beta(M)$ on $[0, 1]$, with $\beta > 0$ and $M > 0$, and*

a) the design $(X_1, \ldots, X_n)$ is regular deterministic;

b) $\mathbb{E}(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2$;

c) $\exists\, \lambda_0 > 0$ such that $\forall\, x \in [0, 1]$, the smallest eigenvalue $\lambda_{\min}(B(x))$ of $B(x)$ satisfies

$$\lambda_{\min}(B(x)) \geqslant \lambda_0 \quad , \textit{ where } B(x) = \frac{1}{n}\sum_{i=1}^{n} U_{x,i}U_{x,i}^T K_h(X_i - x);$$

d) $supp(K) \subseteq [-1, 1]$ and $\exists\, K_{\max} \in (0, \infty)$ such that $\forall u, |K(u)| \leqslant K_{\max}$.

Let $\widehat{f}_n^{LP}$ be the Local Polynomial estimator of f which satisfies the above assumptions with $k = \lfloor \beta \rfloor$. Then, for all $x \in [0,1]$ and $h \geqslant \frac{1}{2n}$,

$$|b(x)| \leqslant \frac{C_K}{k!} M h^{\beta}, \quad v(x) \leqslant \frac{\sigma^2 C_K^2}{nh} \quad \textit{with } C_K = \frac{2K_{\max}}{\lambda_0}.$$

Note that if $\beta \in (0,1)$, the LP estimator becomes the NW estimator, and this proposition coincides with Proposition 10.2.

Now we study consistency and the rates of convergence of $\widehat{f}_n^{LP}(x)$. Under the assumptions of Proposition 10.3, MSE of $\widehat{f}_n^{LP}(x)$ is bounded by

$$\text{MSE}\left[\widehat{f}_n^{LP}(x)\right] = [b(x)]^2 + v(x) \leqslant \left[\frac{C_K}{k!} M\right]^2 h^{2\beta} + \frac{\sigma^2 C_K^2}{n} h^{-1}$$

which is minimised at

$$h = h_n = \left(\frac{\frac{\sigma^2 C_K^2}{n}}{2\beta\left(\frac{C_K M}{k!}\right)^2}\right)^{\frac{1}{2\beta+1}} = \left(\frac{\sigma^2 (k!)^2}{2\beta M^2 n}\right)^{\frac{1}{2\beta+1}},$$

with the value of the minimum being

$$\text{MSE}\left[\widehat{f}_{n,h_{opt}}^{LP}(x)\right] \leqslant \left\{\left[\frac{C_K}{k!} M\right]^2 h_{opt}^{2\beta} + \frac{\sigma^2 C_K^2}{n} h_{opt}^{-1}\right\} = C_{LP} \cdot n^{-\frac{2\beta}{2\beta+1}} \to 0 \text{ as } n \to \infty,$$

where C_{LP} is a constant depending only on M, k, σ^2 and C_K (i.e. $K_{\max}, \lambda_0$).

Now we study the local and global minimax rates of convergence of the LP(k) estimator with $h_n = \alpha n^{-\frac{1}{2\beta+1}}$ over $H^{\beta}(M)$ with $k = \lfloor \beta \rfloor$. In this case, under the conditions of Proposition 10.3,

$$\text{MSE}\left[\widehat{f}_{n,h_n}^{LP}(x)\right] \leqslant C_K^2 \left[\frac{\alpha^2 M^2}{[k!]^2} + \alpha^{-1}\sigma^2\right] n^{-\frac{2\beta}{2\beta+1}},$$

which also implies that

$$\text{MISE }(\widehat{f}^{LP}(x)) = \textstyle\int_0^1 \text{MSE}(\widehat{f}^{LP}(x))dx \leqslant C n^{-\frac{2\beta}{2\beta+1}}$$

with the same constant as in the upper bound on the MSE. Therefore, both local and global rates of convergence of LP(k) are $n^{-\frac{\beta}{1+2\beta}}$. Therefore, the local polynomial estimator achieves both local and global minimax rates of convergence. Hence, we proved the following theorem.

Theorem 10.2. *Under the assumptions of Proposition 10.3, the Local Polynomial estimator with the bandwidth $h = h_n = \alpha n^{-\frac{1}{2\beta+1}}$, $\alpha > 0$, satisfies*

$$\limsup_{n\to\infty} \sup_{f \in \mathbb{H}^{\beta}(M)} \sup_{x_0 \in [0,1]} \mathbb{E}\left[n^{\frac{\beta}{2\beta+1}} |f(x_0) - \widehat{f}_n(x_0)|\right]^2 \leq C < \infty,$$

$$\limsup_{n\to\infty} \sup_{f \in \mathbb{H}^{\beta}(M)} \mathbb{E}\left[n^{\frac{\beta}{2\beta+1}} ||f - \widehat{f}_n||_2\right]^2 \leq C < \infty,$$

where C is a constant depending only on β, M, a_0, λ_0, $\sigma_{\max}^2$, $K_{\max}$ and α.

10.3 Smoothing Splines

10.3.1 Definition

Definition 10.16. *A smoothing spline is the penalised least squares estimator of f:*

$$\widehat{f}_n^{\text{pen}}(x) = \arg\min_{f \in \mathcal{C}^2} \left[\sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \, \text{pen}(f) \right] \tag{125}$$

with penalty function $\text{pen}(f) = \int [f''(x)]^2 dx = ||f''||_2^2$; $\lambda > 0$ *is called the regularisation parameter.*

The solution to this minimisation problem has a simple form that is called a **natural cubic spline**.

Definition 10.17. *Let $a \leq t_1 < .. < t_N \leq b$ be a set of ordered points - called knots. A cubic spline is a continuous function g such that*

- *$g(x)$ is cubic on $[t_j, t_{j+1}]$, for each $j = 1, .., N-1$:*

$$g(x) = b_{j0} + b_{j1}x + b_{j2}x^2 + b_{j3}x^3, \quad x \in [t_j, t_{j+1}],$$

- *both g' and g'' are continuous at t_i, $i = 1, .., N$.*

A spline that is linear beyond the boundary knots is called a natural spline.

- *$g(x)$ is linear on $[a, t_1]$ and $[t_N, b]$*

$$\begin{aligned} g(x) &= b_{00} + \ b_{01}x \ , \quad x \in [a \ , t_1] \\ g(x) &= b_{N0} + b_{N1}x, \quad x \in [t_N, b] \end{aligned}$$

Theorem 10.3. *(without proof) Solution $\widehat{f}_n^{\text{pen}}$ of the above problem is a* **natural cubic spline** *with knots at the data points.*

Theorem 10.4. *Let knots $a \leqslant t_1 < \cdots < t_N \leqslant b$. For $j = 3, \ldots, N$, define*

$$\begin{aligned} h_1(x) &= 1, h_2(x) = x, \\ h_j(x) &= (x - t_{j-2})_+^3 - \frac{(t_N - t_{j-2})}{(t_N - t_{N-1})}(x - t_{N-1})_+^3 \\ &\qquad + \frac{(t_{N-1} - t_{j-2})}{(t_N - t_{N-1})}(x - t_N)_+^3, \quad \forall \ 3 \leqslant j \leqslant N, \\ \textit{where} \qquad & (x-y)_+^3 = \max\left\{(x-y)^3, 0\right\} \end{aligned}$$

The set of functions $(h_j)_{j=1}^N$ forms a basis for the set of natural cubic splines at these knots.

Thus, any natural cubic spline $g(x)$ can be written as

$$g(x) = \sum_{j=1}^{N} \beta_j h_j(x).$$

By Theorem 10.3, the solution of the minimisation problem that defines the smoothing spline is a natural cubic spline, and by Theorem 10.4, it can be written as the linear combination of the basis functions $h_j(x)$, $j = 1, 2, \dots, N$. Hence, minimising over functions f

$$\begin{aligned}\widehat{f}_{n,\lambda}^{SS} &= \arg\min_{f \in C^2} \left\{ \sum_{i=1}^{N} (Y_i - f(x_i))^2 + \lambda \int \left[f''(x)\right]^2 dx \right\} \\ &= \arg\min_{f \in C^2} \left\{ \sum_{i=1}^{N} \left(f(x_i)^2 - 2f(x_i)Y_i + Y_i^2\right) + \lambda \int \left[f''(x)\right]^2 dx \right\}\end{aligned}$$

is equivalent to minimising the following expression over the $(n+2)$-dimensional vector β:

$$\begin{aligned}\widehat{\beta} &= \arg\min_{\beta \in \mathbb{R}^N} \left\{ \sum_{i=1}^{N} \left[\sum_{j=1}^{N} \beta_j h_j(x_i)\right]^2 - 2\sum_{i=1}^{N} \left[\sum_{j=1}^{N} \beta_j h_j(x_i)\right] Y_i + \lambda \int \left[\sum_{j=1}^{N} \beta_j h_j''(x)\right]^2 dx \right\} \\ &= \arg\min_{\beta \in \mathbb{R}^N} \left\{ \beta^T H^T H \beta - 2\beta^T H^T Y + \lambda \beta^T \Omega \beta \right\},\end{aligned}$$

where $N \times N$ matrix H has entries $H_{ij} = h_j(x_i)$, $i = 1, \dots, N$, $j = 1, \dots, N$, and $N \times N$ matrix Ω has elements $\Omega_{j\ell} = \int h_j''(x) h_\ell''(x) dx$, $j, \ell = 1, \dots, N$.

Hence, if $\left(H^T H + \lambda \Omega\right)$ is invertible,

$$\widehat{\beta} = \left[\left(H^T H + \lambda \Omega\right)^{-1} H^T Y\right].$$

Therefore, we have proved the following theorem.

Theorem 10.5. *A smoothing spline can be written as*

$$\widehat{f}_{n,\lambda}^{SS} = \sum_{j=1}^{N} \widehat{\beta}_j h_j(x)$$

where $\widehat{\beta} = (\widehat{\beta}_1, \dots, \widehat{\beta}_N)^T$ *is given by*

$$\widehat{\beta} = (H^T H + \lambda \Omega)^{-1} H^T Y$$

where $Y = (Y_1, \dots, Y_n)^T$*, and matrices* $H = (H_{ij})$ *and* $\Omega = (\Omega_{jl})$ *have entries*

$$H_{ij} = h_j(x_i), \quad \Omega_{jl} = \int_a^b h_j''(x) h_l''(x) dx, \quad i \in 1, \dots, n, \quad j, l \in 1, \dots, N$$

The smoothing spline is a linear estimator since it can be written as

$$\widehat{f}_{N,\lambda}^{SS} = \sum_{i=1}^{N} w_i(x) Y_i$$

with weights

$$w_i(x) = \sum_{j=1}^{N} h_j(x) \left[\left(H^T H + \lambda \Omega\right)^{-1} H^T\right]_{ji}$$

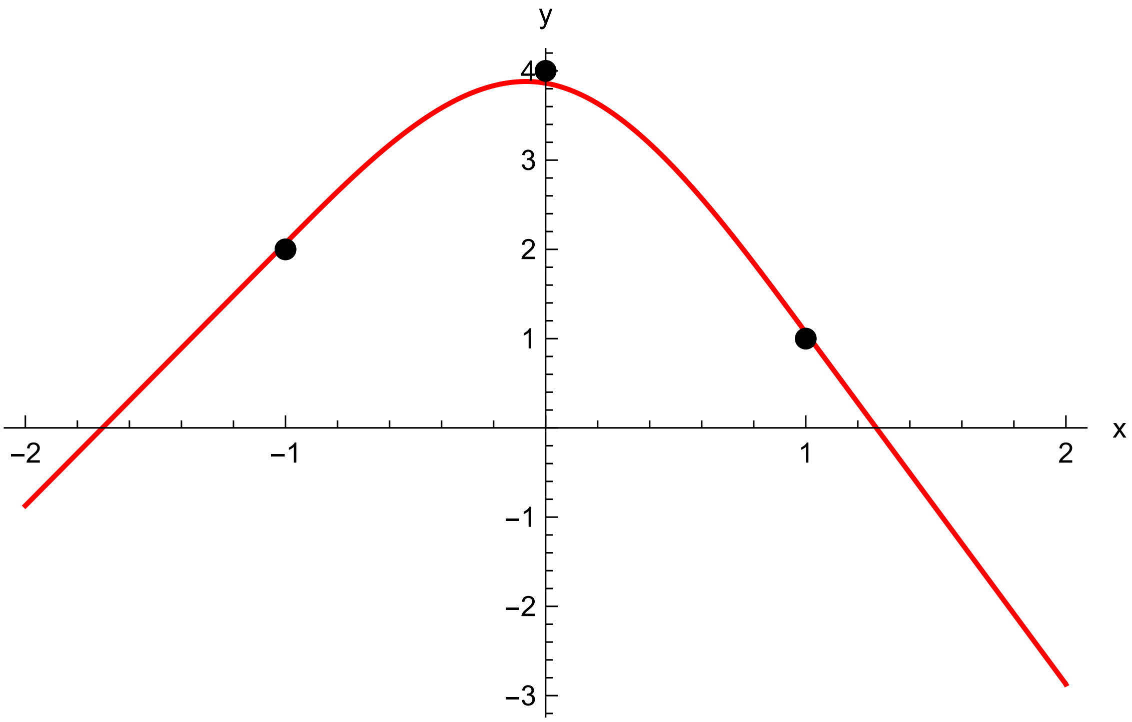


Figure 42: Smoothing spline for example 10.7.

Example 10.7. *Construct a smoothing spline on* $[-2, 2]$ *given data* $(-1, 2)$, $(0, 4)$, $(1, 1)$. *Take* $\lambda = 0.01$, *and construct the smoothing spline using*

$$\widehat{f}_n^{SS}(x) = \sum_{i=1}^{N}\sum_{j=1}^{N}[(H^T H + \lambda\Omega)^{-1}H^T]_{ji}h_j(x)Y_i.$$

The matrices necessary for the calculation are $H = (H_{ij})$, $H_{ij} = h_j(x_i)$*:*

$$H = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 6 \end{pmatrix}, \quad H^T H = \begin{pmatrix} 3 & 0 & 7 \\ 0 & 2 & 6 \\ 7 & 6 & 37 \end{pmatrix}$$

and $\Omega = (\Omega_{j\ell})$, $\Omega_{j\ell} = \int h_j''(x)h_\ell''(x)dx$*:*

$$\Omega = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 24 \end{pmatrix}$$

We find the coefficients of the natural spline are $\hat{\beta}^T = (5.00917, 2.94037, -1.14679)$. *The data and smoothing spline are shown in Figure 42.*

10.3.2 Choice of Regularisation Parameter $\boldsymbol{\lambda}$

In papplications, λ is usually chosen using cross-validation

$$\widehat{\lambda} = \arg\min_{\lambda > 0}\left\{\sum_{i=1}^{n}\left(Y_i - \widehat{f}_{\lambda,-i}(x_i)\right)^2\right\}$$

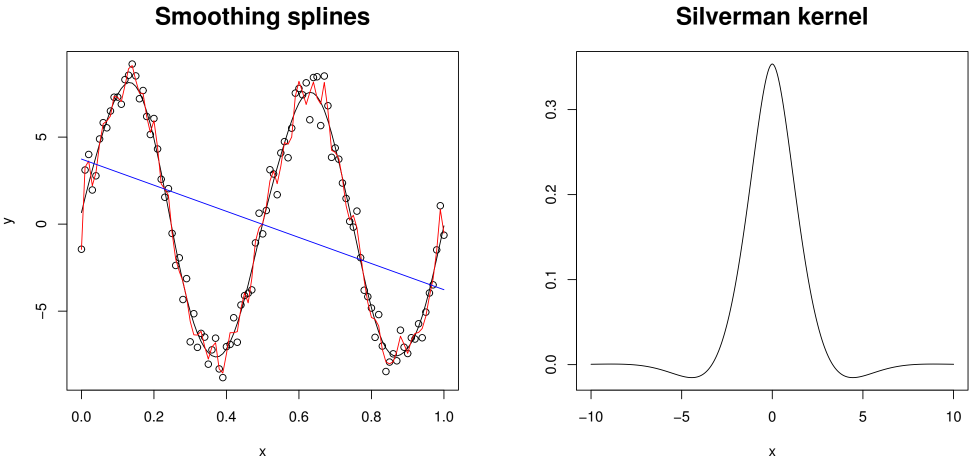


Figure 43: Left: smoothing spline estimator Right: Silverman kernel

where $\widehat{f}_{\lambda,-i}$ is a smoothing spline based on all data points except the i'th. The expression to be minimised is an unbiased estimator of MISE.

Smoothing spline estimators with different regularisation parameters λ are plotted in Figure 43 (Left). The black line corresponds to λ is chosen by cross-validation, the red line - to $\lambda = 0.05$, and the blue line - to $\lambda = 2$. For small $\lambda = 0.05$, where the leading contribution comes from the likelihood, the fitted curve is close to the data points but is not particularly smooth. For larger $\lambda = 2$, the penalisation term dominates the likelihood term, and the linear curve is such that the penalty term is zero (since the second derivative of a linear function is 0). λ chosen by cross-validation provides the estimator with the trade-off between fit to the observed data and smoothness.

10.3.3 Smoothing Spline as a Kernel Estimator

For large N, the smoothing spline is asymptotically equivalent to a kernel estimator:

$$\widehat{f}^{SS}(x) \approx \widehat{f}^{NW}(x),$$

where $\widehat{f}^{NW}(x)$ is the Nadaraya-Watson estimator with the Silverman kernel:

$$K(z) = \frac{1}{2} e^{-|z|/\sqrt{2}} \sin(|z|/\sqrt{2} + \pi/4),$$

plotted in Figure 43 (right), and the bandwidth h can be expressed in terms of λ as $h = \lambda^{1/4}$. Note that this kernel can take negative values. In particular, the smoothing spline has the same optimality properties as a kernel estimator, such as consistency and the optimal rates of convergence.

10.4 Generalized Additive Models

So far we have only talked about regression models with one covariate. However, a more common regression problem would have multiple covariates and take the form

$$Y_i = f(x_{1i}, x_{2i}, \ldots, x_{mi}) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $x_1, \ldots, x_m$ are a set of covariates. Fitting of multivariate regression models is more challenging, not least because large amounts of data are in general required to ensure convergence. The optimal rate of convergence for $f \in H^2(M)$ (i.e., functions with an integrable second derivative) is $n^{-4/5}$ with one covariate, but this degrades to $n^{-4/(4+m)}$ when there are m covariates. If n is the sample size required to achieve a certain accuracy with one covariate, then the sample size required to achieve the same accuracy with m covariates is $n^{(4+m)/5}$ and therefore grows exponentially with m. Nonetheless, generalisations of most univariate nonparametric methods exist and we will describe some of these here.

10.4.1 Multivariate local polynomial regression

Kernel regression can be carried out with multiple covariates, but requires generalisation of the kernel function so that it is a function of m variables. The one-dimensional bandwidth h is replaced by a bandwidth matrix H, allowing a family of kernels to be defined via

$$K_H(\mathbf{x}) = \frac{1}{\sqrt{\det(H)}} K\left(H^{-1/2}\mathbf{x}\right).$$

A common approach is to rescale the covariates so that they have the same mean and variance (at least approximately) and then use an isotropic kernel $h^{-m}K(||\mathbf{x}||_2/h)$ where $K(\cdot)$ is a one-dimensional kernel.

Given a choice of kernel, the local polynomial estimator of order k is found in the same way as before. Firstly we note that an arbitrary function of m variables can be expanded as

$$\begin{aligned} f(x_1, \ldots, x_m) = f(\mathbf{z}) &+ \frac{\partial f}{\partial x_1}(x)(x_1 - z_1) + \frac{\partial f}{\partial x_2}(x)(x_2 - z_2) + \cdots + \frac{\partial f}{\partial x_m}(x)(x_m - z_m) \\ &+ \frac{1}{2!}\left(\frac{\partial^2 f}{\partial x_1^2}(x)(x_1 - z_1)^2 + 2\frac{\partial^2 f}{\partial x_1 \partial x_2}(x)(x_1 - z_1)(x_2 - z_2) + \right. \\ &\qquad \left. \cdots + \frac{\partial^2 f}{\partial x_m^2}(x)(x_m - z_m)^2\right) + \cdots \\ &+ \frac{1}{k!}\left(\frac{\partial^k f}{\partial x_1^k}(x)(x_1 - z_1)^k + \cdots + \frac{\partial^k f}{\partial x_m^k}(x)(x_m - z_m)\right). \end{aligned}$$

There are a total of $M_k = {}_{m+k}C_m = (m+k)!/(m!k!)$ distinct partial derivative terms in this expansion. We can define analogues of the parameter vector θ and the design vector $U_{x,i}$ with this many components

$$\theta_{\mathbf{x}} = (\theta^0, \theta^1_1, \theta^1_2, \ldots, \theta^1_m, \theta^2_{11}, \theta^2_{12}, \ldots \theta^2_{mm}, \ldots, \theta^k_{mm\cdots m})$$

$$\begin{aligned} U_{\mathbf{x},i} = \Bigg(1, \frac{x_{1i} - x_1}{h}, \frac{x_{2i} - x_2}{h}, \ldots, \frac{x_{mi} - x_m}{h}, \frac{1}{2!}\left(\frac{x_{1i} - x_1}{h}\right)^2, \left(\frac{x_{1i} - x_1}{h}\right)\left(\frac{x_{2i} - x_2}{h}\right), \\ \ldots, \frac{1}{2!}\left(\frac{x_{mi} - x_m}{h}\right)^2, \ldots, \frac{1}{k!}\left(\frac{x_{mi} - x_m}{h}\right)^k\Bigg). \end{aligned}$$

In the above, $h^m = \sqrt{\det(H)}$, $\theta^d_{j_1 \dots j_d}$ corresponds to $h^d \partial^d f / \partial x_{j_1} \cdots \partial x_{j_d}$ and the estimator of this quantity provides an estimate of this particular derivative of the function. Note that we must be careful to ensure the ordering of derivatives in θ and $U_{\mathbf{x},i}$ is consistent.

Using this notation the solution for the local polynomial least squares estimator

$$\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^{M_d}} \left\{ \sum_{i=1}^{n} \left(Y_i - U^T_{\mathbf{x},i} \theta_{\mathbf{x}} \right)^2 K_H(\mathbf{x}_i - \mathbf{x}) \right\}$$

takes exactly the same form as before, namely $\hat{\theta}_{\mathbf{x}} = B^{-1}(\mathbf{x}) a(\mathbf{x})$ where

$$B(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} U_{\mathbf{x},i} U^T_{\mathbf{x},i} K_H(\mathbf{x}_i - \mathbf{x}), \qquad a(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} Y_i U_{\mathbf{x},i} K_H(\mathbf{x}_i - \mathbf{x}).$$

10.4.2 Multivariate splines

In a similar way, the notion of a spline can be generalized to more than one dimension. Once again, we aim to minimize the sum of squares, but penalise functions that are not sufficiently smooth. This is formulated in general as

$$\hat{f}^{SS}_{n,\lambda} = \arg \min_f \left\{ \sum_{i=1}^{n} (Y_i - f(x_{1i}, \dots, x_{mi}))^2 + \lambda J_n(f) \right\}$$

where

$$J_n(f) = \int \int \cdots \int \left[\left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_1 \partial x_2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_1 \partial x_3} \right)^2 + \right.$$
$$\left. \cdots + \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_2 \partial x_3} \right)^2 + \cdots \left(\frac{\partial^2 f}{\partial x_m^2} \right)^2 \right] \mathrm{d}x_1 \mathrm{d}x_2 \dots \mathrm{d}x_m.$$

The solution to the minimization problem is a **thin plate spline**.

Definition 10.18. *A* **thin plate spline** *through a set of knots $\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_n$ in m-dimensions, with weights $w_1, \dots w_n$, is a function of the form*

$$f(\mathbf{x}) = \sum_{i=1}^{n} w_i G(||\mathbf{x} - \mathbf{x}_i||_2) + b_0 + \sum_{j=1}^{m} b_j x_j$$

$$\text{where } G(r) \propto \begin{cases} r^{4-m} \ln r, & m = 2 \text{ or } m = 4 \\ r^{4-m}, & \text{otherwise} \end{cases}, \qquad \text{and } ||\mathbf{x}||_2^2 = \sum_{j=1}^{m} x_j^2.$$

In higher dimensions, $m > 4$, this solution diverges at the knots and so it is not a useful smoothing method. In that case the $m = 2$ basis function, $G(r) = r^2 \ln r$, is often used, or the simple solution $G(r) = r^2$. If these alternative solutions are used the resulting solution is in general not the minimizer for the above problem.

Thin plate splines are difficult to fit and so are not used widely in dimensions higher than 2. It is more common to take an approach that reduces the multi-dimensional fit to a set of one-dimensional fits by using an **additive model**.

10.4.3 Additive models

While the preceding methods provide ways to fit general multivariate nonparametric models, they are often hard to visualize and interpret. This motivates assuming a somewhat simpler form for the unknown function, called an **additive model**.

Definition 10.19. *An additive model is a model of the form*

$$Y_i = \alpha + \sum_{j=1}^{m} f_j(x_j) + \epsilon_i, \qquad i = 1, \ldots, n$$

where $f_1, \ldots, f_m$ are smooth functions.

The model above is not identifiable since a constant can be subtracted from any one of the functions and added to α or any of the other functions to leave the model unchanged. The usual approach to making the model identifiable is to set $\hat{\alpha} = \bar{Y} = \sum_{i=1}^{n} Y_i/n$ and forcing $\sum_{i=1}^{n} \hat{f}_j(x_{ji}) = 0$. The resulting functions can be regarded as representing deviations from the mean $\bar{Y}$.

An additive model can be fitted using any of the techniques for one-dimensional problems that have been described in this course using a procedure known as **backfitting**.

Definition 10.20. *The backfitting algorithm obtains estimates of $\hat{f}_j(x_j)$ in the additive model as follows. Fix the estimator $\hat{\alpha} = \bar{Y}$ and choose initial guesses for $\hat{f}_1, \ldots, \hat{f}_m$. Then*

1. *For $j = 1, \ldots, m$:*
   - *(a) Compute $\tilde{Y}_i = Y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ki}), i = 1, \ldots, n$.*
   - *(b) Apply a one-dimensional nonparametric fitting procedure (smoother) to $\tilde{Y}_i$ as a function of x_j. Set $\hat{f}_j$ equal to the output of this procedure.*
   - *(c) Renormalise by setting $\hat{f}_j(x)$ equal to $\hat{f}_j(x) - \sum_{i=1}^{n} \hat{f}_j(x_{ji})/n$.*
2. *Repeat step 1 until the estimators converge.*

10.4.4 Projection pursuit

Projection pursuit regression attempts to approximate the unknown function $f(x_1, \ldots, x_m)$ by one of the form

$$\mu + \sum_{j=1}^{M} r_j(z_j) \qquad \text{where } z_i = \alpha_i^T \mathbf{x}$$

and each α_i is a unit vector. Projection pursuit attempts to find a transformation of the coordinates that makes an additive model fit as well as possible. In practice, projection pursuit is fitted iteratively, using some one-dimensional nonparametric method. We use $S(w; \mathbf{Y}, \mathbf{x})$ to denote the value of the output of this nonparametric method at a point w, where $\mathbf{x}$ is the vector of (one-dimensional) covariates at the observed points and $\mathbf{Y}$ is the vector of measured values. First set $\hat{\mu} = \bar{Y}$ as before and then initialise the residuals $\hat{\epsilon}_i = Y_i - \bar{Y}$. We use $\hat{\epsilon}$ to denote the vector of current residuals, i.e., $(\hat{\epsilon})_i = \hat{\epsilon}_i$. We also scale the covariates so that their variances are equal and then define an $m \times n$ matrix X such that X_{ij} is the value of the i'th covariate for the j'th data point. Then proceed as follows:

1. Set $j = 0$.

2. Find the unit vector α that minimizes

$$I(\alpha) = 1 - \frac{\sum_{i=1}^{n}(\hat{\epsilon}_i - S(\alpha^T \mathbf{x}_i; \hat{\epsilon}, X^T\alpha))^2}{\sum_{i=1}^{n} \hat{\epsilon}_i^2}$$

   and then set $z_{ji} = \alpha^T \mathbf{x}_i$ and $\hat{f}_j(z_{ji}) = S(\alpha^T \mathbf{x}_i; \hat{\epsilon}, X^T\alpha)$.

3. Set $j = j + 1$ and update the residuals

$$\hat{\epsilon}_i \leftarrow \hat{\epsilon}_i - \hat{f}_j(z_{ji}).$$

4. If $j = M$ stop, else return to step 2.

10.4.5 Generalized additive models

Definition 10.21. *An generalized additive model is a model in which observed random variables Y_i are assumed to be drawn from a specified distribution in the exponential family, with a specified link function, $g(\cdot)$, and a model for the expectation value of the form*

$$\eta(\mathbf{x}) = g(\mathbb{E}(Y)) = \alpha + \sum_{j=1}^{m} f_j(x_j)$$

where $f_1, \ldots, f_m$ are smooth functions.

Fitting a generalized additive model can be done iteratively, using a method for fitting a general additive model, in the same way that generalized linear models can be found by fitting general linear models using iterative weighted least squares (Fisher's method of scoring).

The general procedure is as follows:

1. Start with observed data $\{(\mathbf{x}_i, y_i) : i = 1, \ldots n\}$ and initial guesses for $\hat{\alpha}$ and $\hat{f}_1, \ldots, \hat{f}_m$.

2. Then repeat the folliwng steps until the estimates for $\hat{f}_1, \ldots, \hat{f}_m$ converge:

   (a) Compute fitted values

$$\hat{\eta}(\mathbf{x}_i) = \hat{\alpha} + \sum_{j=1}^{m} \hat{f}_j(x_{mi})$$
$$\text{and } \hat{r}(\mathbf{x}_i) = g^{-1}(\hat{\eta}(\mathbf{x}_i)).$$

   (b) Computed transformed responses

$$z_i = \hat{\eta}(\mathbf{x}_i) + (y_i - \hat{r}(\mathbf{x}_i))g'(\hat{r}(\mathbf{x}_i)),$$

   where $g'(\cdot)$ denotes the derivative of the link function.

   (c) Compute weights

$$w_i = \left[(g'(\hat{r}(\mathbf{x}_i))^2\sigma^2\right]^{-1}.$$

(d) Compute the weighted general additive model for z_i as a function of $\mathbf{x}_i$ with weights w_i.

Note that the above procedure relies on being able to fit a weighted nonparametric model, but all of the methods described above have assumed equal variance. However, it is straightforward to generalise the previous methods to the weighted context. For example, the extension of the Nadaraya-Watson estimator to the weighted case is

$$\hat{f}_n^{wNW}(x) = \frac{\sum_{i=1}^n w_i Y_i K_h(X_i - x)}{\sum_{j=1}^n w_j K_h(X_j - x)}.$$

Example 10.8. *Construct a general additive model, using smoothing splines, on the interval* $[-2,2] \times [-2,2]$ *given data* $(-1,-1,1)$, $(-1,0,3)$, $(-1,1,0)$, $(0,-1,2)$, $(0,0,4)$, $(0,1,1)$, $(1,-1,6)$, $(1,0,3)$, $(1,1,2)$. *Use* $\lambda = 0.01$ *in both dimensions.*

We note that in this case we have data on a regular grid. The backfitting procedure fits a function in one dimension at a time, and so we will need to fit a smoothing spline with multiple observations at a given point. For equal numbers of observations at each point, n_s, *this is a trivial extension of the procedure described above. The spline takes the same form, but we replace* Y_i *by the average of the* Y_i'*s at each value of* x, *and we change the smoothing parameter to* λ/n_s.

First we estimate $\hat{\alpha} = \bar{Y} = 22/9$ *and subtract this from each point. We then fit a smoothing spline to the data* $(-1,-10/9)$, $(0,-1/9)$, $(1,11/9)$ *using* $\lambda = 0.01/3$. *The* H *and* Ω *matrices are the same as in Example* 3.1

$$H = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 6 \end{pmatrix}, \qquad \Omega = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 24 \end{pmatrix}.$$

and we derive $\hat{\beta}_1 = [(H^T H + \lambda\Omega)^{-1} H^T Y]$ *as before*

$$\hat{\beta}_1^T = (-0.188781, 0.923948, 0.0809061).$$

This gives fitted values at $x = -1, 0, 1$ *of*

$$\hat{f}_1(-1) = -1.11273, \quad \hat{f}_1(0) = -0.107875, \quad \hat{f}_1(1) = 1.2206.$$

We need to correct the fit by subtracting $\sum_{i=1}^3 \hat{f}_1(x_{1i})/3$, *but this number is very close to zero so the values do not change.*

We now need to fit for the second dimension, x_2. *The first stage, in general, is to subtract* $\hat{f}_1(x_{1i})$ *from* Y_i *for each* i. *In this case we have multiple observations at each value of* x_2 *and so we then need to average the* Y_i*'s for each* x_2. *Since the grid is regular, we effectively subtract* $\sum_{i=1}^3 \hat{f}_1(x_{1i})/3$ *from each value, but this has been fixed to equal* 0 *and so does not change the averaged values. This happens generically when the data is on a regular grid and means the backfitting algorithm converges in one iteration.*

The data to fit in x_2 *is* $(-1,5/9)$, $(0,8/9)$, $(1,-13/9)$ *with* $\lambda = 0.01/3$ *again. The* H *and* Ω *matrices are unchanged so we obtain*

$$\hat{\beta}_2^T = (1.51025, 0.941748, -0.647249).$$

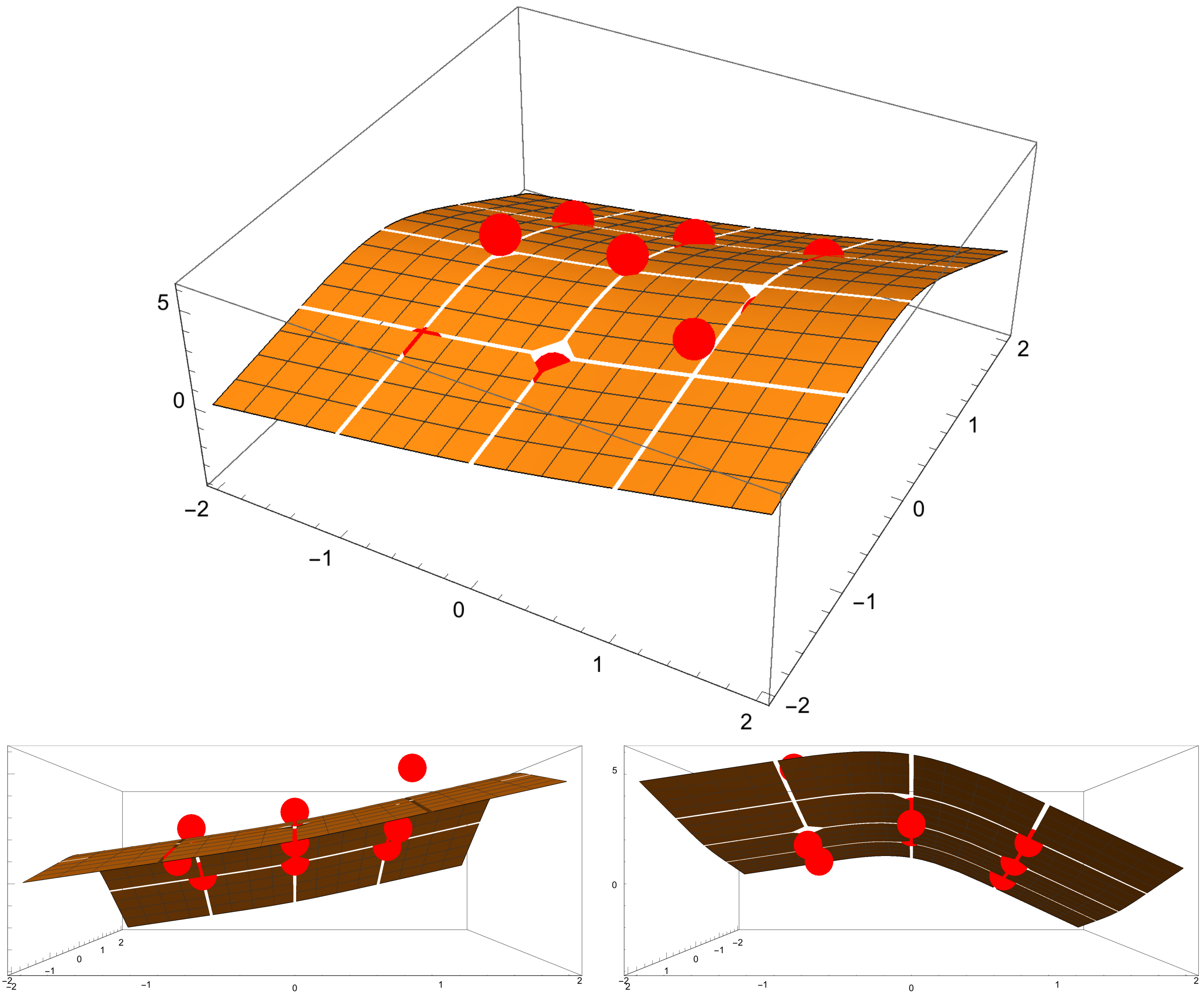


Figure 44: Data (red points) and general additive model fit (shaded surface) for example 10.8. The top plot shows the full surface, while the bottom two plots show the surface from the x_1 and x_2 sides respectively.

The algorithm has now converged and we obtain our general additive model estimate of $f(x_1, x_2)$ *as*

$$\hat{f}(x_1, x_2) = \frac{22}{9} + \sum_{i=1}^{3} \beta_{1i} h_i(x_1) + \sum_{i=1}^{3} \beta_{2i} h_i(x_2)$$

$$\text{where } h_1(x) = 1, \qquad h_2(x) = x, \qquad h_3(x) = (x+1)_+^3 - 2(x)_+^3 + (x-1)_+^3.$$

The raw data and the GAM estimate are shown in Figure 44.

10.5 Wavelet Estimators

We return again to the nonparametric regression model

$$Y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \dots, n, \qquad \mathbb{E}(\varepsilon_i) = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma^2,\ \text{independently.}$$

In this subsection we will assume that the design is regular deterministic, that is $x_i - x_{i-1} = 1/n$ for all i. In particular, we consider $x_i = \frac{i}{n}$.

10.5.1 Orthonormal basis and projection estimator

We will denote the set of square-integrable functions by $L^2 = \left\{ f : ||f||_2 = \sqrt{\int f^2(x)dx} < \infty \right\}$.

Definition 10.22. *A set of functions $\{\varphi_k(x)\}_{k=0}^{\infty}$ is called an orthonormal basis of $L^2[0,1]$, if*

- $\forall\, g \in L^2, \exists\, (a_k)_{k=0}^{\infty}$ *such that* $g(x) = \sum_{k=0}^{\infty} a_k \varphi_k(x)$ *(the set spans $L^2[0,1]$),*
- $\forall\, x, \sum_{k=0}^{\infty} a_k \varphi_k(x) = 0 \Rightarrow$ *all* $a_k = 0$ *(linear independence),*
- $j \neq k, \int \varphi_k(x)\varphi_j(x) = 0$ *(orthogonality),*
- $\forall\, k, ||\varphi_k||_2 = 1$ *(normalisation).*

Therefore, any function $f \in L^2[0,1]$ can be written as

$$f(x) = \sum_{k=0}^{\infty} \theta_k \varphi_k(x).$$

Due to orthonormality of the basis, the coefficients θ_k have a simple expression: $\theta_k = \int_0^1 f(x)\varphi_k(x)dx$, since

$$\int_0^1 f(x)\varphi_k(x)dx = \int_0^1 \left[\sum_{j=0}^{\infty} \theta_j \varphi_j(x) \right] \varphi_k(x)dx = \sum_{j=0}^{\infty} \theta_j \left[\int_0^1 \varphi_j(x)\varphi_k(x)dx \right] = \theta_k$$

Examples of orthonormal bases:

1. Fourier basis: $\varphi_{2k}(x) = 1$, $\varphi_{2k}(x) = \cos(2\pi kx)$, $\varphi_{2k+1}(x) = \sin(2\pi kx)$, $k = 1, 2, \dots$, $x \in [0,1]$ (Tsybakov, 2009).
2. A wavelet basis (Vidakovic, 1999)
3. An orthogonal polynomial basis, such as Chebyshev, Lagrange, Laguerre polynomials (more commonly used in the context of density estimation)

Projection estimator

Assume that $f \in L^2[0,1]$, and $\{\varphi_k(x)\}_{k=0}^{\infty}$ is an orthonormal basis of $L^2[0,1]$. Then, we can write

$$f(x) = \sum_{k=0}^{\infty} \theta_k \varphi_k(x)$$

for some real coefficients $\theta_0, \theta_1, \ldots$. A projection estimation of f is based on a simple idea: approximate f by its projection $\sum_{k=0}^{N} \theta_k \varphi_k(x)$ on the linear span of the first $N+1$ functions of the basis, and replace θ_k by their estimators. Thus, a projection estimator is constructed in three steps.

$$\text{(1)} \qquad \text{for large } N, \text{ approximate } f(x) \approx \sum_{k=0}^{N} \theta_k \varphi_k(x)$$

$$\text{(2)} \qquad \text{construct an estimator } \widehat{\theta}_k \text{ of } \theta_k \text{ from data } (y_1, \ldots, y_n),\ k = 0, 1, \ldots, N$$

$$\text{(3)} \qquad \text{plug in the estimator } \widehat{\theta}_k \text{ in the approximation: } \widehat{f_N}(x) = \sum_{k=0}^{N} \widehat{\theta}_k \varphi_k(x)$$

From the expression for θ_k in terms of f and φ_k, if we know only values of $f(x)$ at points $x_i = i/n$, $i = 1, \ldots, n$, then for large n the integral can be approximated by a sum:

$$\theta_k \approx \frac{1}{n} \sum_{i=1}^{n} f(x_i) \varphi_k(x_i).$$

Since we observe values of $f(x_i)$ with error, we plug in these observation in the above expression to obtain the following estimator for θ_k:

$$\widehat{\theta}_k = \frac{1}{n} \sum_{i=1}^{n} Y_i \varphi_k(x_i).$$

Inserting this expression into the estimator of the function, we obtain a **projection estimator**:

$$\widehat{f}_N(x) = \sum_{k=0}^{N} \left[\frac{1}{n} \sum_{i=1}^{n} f(x_i) \varphi_k(x_i) \right] \varphi_k(x) = \sum_{i=1}^{n} Y_i \left[\sum_{k=0}^{N} \frac{1}{n} \varphi_k(x_i) \varphi_k(x) \right]$$

which is a linear estimator with weights $w_i(x) = \sum_{k=0}^{N} \frac{1}{n} \varphi_k(x_i) \varphi_k(x)$ which do not depend on Y_i. The choice of N corresponds to choosing the smoothness of the function $\widehat{f}_N$.

10.5.2 Wavelet basis

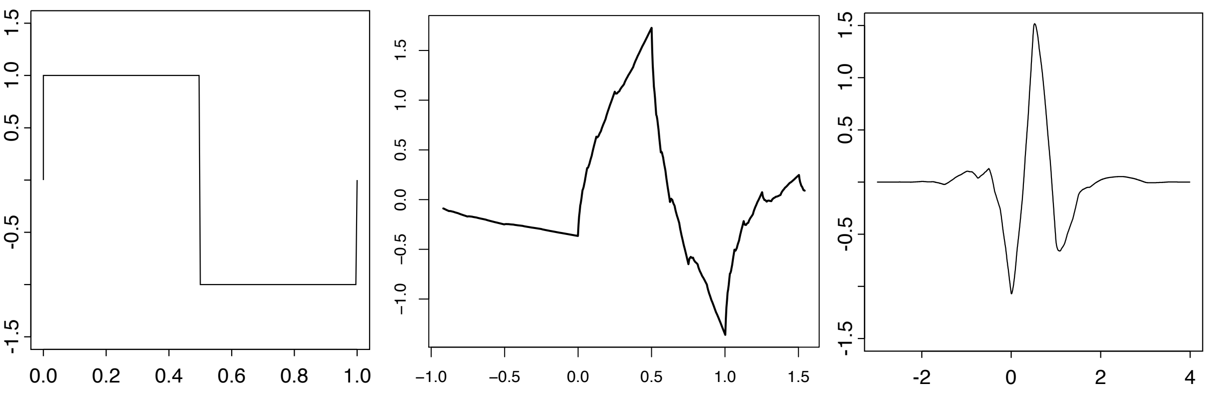
A wavelet basis is constructed using two functions, a scaling function $\phi(x)$ and a wavelet function $\psi(x)$ that are also called the father and mother wavelet respectively. They satisfy the following properties:

$$\int \phi(x) dx = 1, \qquad \int \psi(x) dx = 0.$$

Definition 10.23. *Given a wavelet function ψ and a scaling function ϕ, a wavelet basis on $[0, 1]$ is*

$$\{\phi,\ \psi_{jk},\ j = 0, 1, \ldots, k = 0, \ldots, 2^j - 1\},$$

where $\phi_{jk}(x) = 2^{j/2} \phi(2^j x - k)$, $\psi_{jk}(x) = 2^{j/2} \psi(2^j x - k)$.



(a) Haar mother wavelet (b) Daubechies mother wavelet, $s = 2$ (c) Daubechies wavelet, $s = 4$

Figure 45: Haar and Daubechies wavelet functions

Under certain additional conditions on the scaling function $\phi(x)$ and the wavelet function $\psi(x)$, this basis is *orthonormal*. Then, any $f \in L^2[0,1]$ can be decomposed in a **wavelet basis**:

$$f(x) = \theta_0 \phi(x) + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \theta_{jk} \psi_{jk}(x),$$

and $\theta = \{\theta_0, \theta_{jk}\}$ is a set of **wavelet coefficients**:

$$\theta_0 = \int_0^1 \phi(x) f(x) dx, \quad \theta_{jk} = \int_0^1 \psi_{jk}(x) f(x) dx.$$

Wavelets (ϕ, ψ) are said to have regularity s if they have s derivatives and ψ has s vanishing moments ($\int x^k \psi(x) dx = 0$ for integer $k \leq s$).

Examples of wavelet functions are plotted in Figure 45, and the structure of the wavelet basis is illustrated in Figure 46.

Example 10.9. *The Haar wavelet basis is determined by the scaling function $\phi(x) = \mathbf{1}_{(0,1]}(x)$ and the wavelet function $\psi(x) = \mathbf{1}_{(0,1/2]}(x) - \mathbf{1}_{(1/2,1]}(x)$ which satisfy*

$$\int \phi(x) dx = 1, \quad \int \psi(x) dx = 0, \quad \int \psi_{jk}(x) dx = 0.$$

Check that the basis $\{\phi,\ \psi_{jk},\ j = 0, 1, \ldots, k = 0, \ldots, 2^j - 1\}$ defined by these functions is orthonormal, that is, that the functions are normalised

$$||\phi||_2^2 = \int \phi^2(x) dx = 1, \qquad ||\psi||_2^2 = \int \psi^2(x) dx = 1, \qquad ||\psi_{jk}||_2^2 = \int \psi_{jk}^2(x) dx = 1,$$

and are orthogonal:

$$\int \phi(x) \psi_{jk}(x) dx = 0, \qquad \int \psi_{jk}(x) \psi_{\ell m}(x) = 0 \ \textit{for}\ (j,k) \neq (\ell, m).$$

Local polynomial and kernel estimators provide localisation in time only. A Fourier basis provides localisation in frequency only. The advantage of a wavelet basis is that it provides localisation in both time and frequency, at the expense of having two indices. The wavelet transform provides a sparse representation of most functions (it is the basis of JPEG2000).

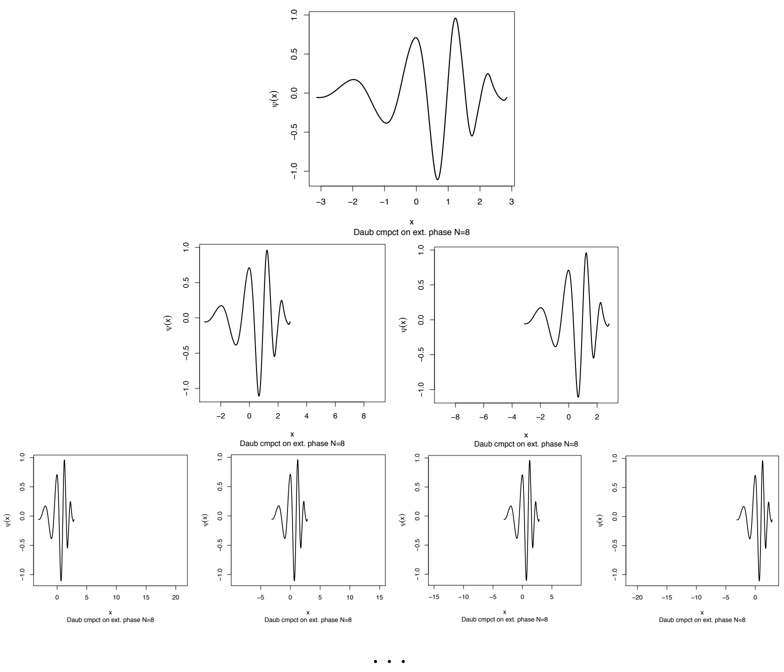


Figure 46: Daubechies wavelet transform, $s = 8$

10.5.3 Wavelet estimators

A **wavelet estimator** can be constructed following the same structure as a projection estimator:

1) derive an estimate $\widehat{\theta}_{jk}$ from noisy discrete wavelet coefficients

2) substitute into the series expansion to obtain the estimate of f, to obtain a wavelet estimator $\widehat{f}$:

$$\widehat{f}(x) = \widehat{\theta}_0 \phi(x) + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \widehat{\theta}_{jk} \psi_{jk}(x).$$

For example, a **wavelet projection estimator** can be constructed as

$$\widehat{f}_{J_0}(x) = \widehat{\theta}_0 \phi(x) + \sum_{j=0}^{J_0-1} \sum_{k=0}^{2^j-1} \widehat{\theta}_{jk} \psi_{jk}(x),$$

with

$$\widehat{\theta}_0 = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi(x_i), \quad \widehat{\theta}_{jk} = \frac{1}{n} \sum_{i=1}^{n} Y_i \psi_{jk}(x_i), \; j < J_0.$$

From this definition it follows that $\widehat{\theta}_{jk} = 0$ for $j \geq J_0$. It is a linear estimator.

The number of nonzero coefficients of $\widehat{f}_{J_0}(x)$ is

$$1 + \sum_{j=0}^{J_0-1} \sum_{k=0}^{2^j-1} 1 = 1 + \sum_{j=0}^{J_0-1} 2^j = 1 + \frac{2^{J_0} - 1}{2 - 1} = 2^{J_0}.$$

Example 10.10. *For the Haar wavelet projection estimator, the variance is*

$$Var(\widehat{f_{J_0}}(x)) = \frac{\sigma^2}{n} \left[(\phi(x))^2 + \sum_{j=0}^{J_0-1} \sum_{k=0}^{2^j-1} (\psi_{jk}(x))^2 \right] = \frac{\sigma^2}{n} \left[1 + \sum_{j=0}^{J_0-1} 2^j \right] = \frac{2^{J_0}}{n} \sigma^2,$$

since $(\phi(x))^2 = 1$ *for all* $x \in [0,1]$*, and* $(\psi_{jk}(x))^2 = 2^j$ *for* (j,k) *such that* $x \in supp(\psi_{jk})$*, i.e. if* $\frac{k}{2^j} \le x < \frac{k+1}{2^j}$ *(just one* $k = \lfloor x2^j \rfloor$ *for each* j *satisfies this condition).*

We will also consider wavelet thresholding estimators which are examples of nonlinear estimators (see Section 10.5.10).

10.5.4 Multiresolution analysis (MRA)

In this section there is a brief explanation of why wavelet functions, together with the scaling function, form a basis.

Definition 10.24. *A multiresolution analysis (MRA) is a sequence of closed subspaces* V_n*,* $n \in \{0,1,2,..\}$ *in* $L^2(\mathbb{R})$ *such that*

1. $V_0 \subset V_1 \subset V_2 \subset \dots, \quad \mathrm{Clos}(\bigcup_j V_j) = L^2(\mathbb{R})$*, where* $\mathrm{Clos}(A)$ *stands for the closure of a set* A*.*

2. *Subspaces* V_j *are self-similar:*

$$g(2^j x) \in V_j \quad \Leftrightarrow \quad g(x) \in V_0,$$

3. *There exists a scaling function* $\phi \in V_0$ *such that* $\int_{\mathbb{R}} \phi(x)dx \neq 0$ *whose integer-translates span the space* V_0*:*

$$V_0 = \left\{ g \in L^2(\mathbb{R}) : \quad g(x) = \sum_{k\in\mathbb{Z}} c_k \phi(x-k) \textit{ for some } (c_k)_{k\in\mathbb{Z}} \right\},$$

and for which the set of functions $\{\phi(\cdot - k), \quad k \in \mathbb{Z}\}$ *is an orthonormal basis.*

Property 2 of MRA implies that for any $h(x) \in V_j$ $\exists\, g \in V_0$ such that

$$h(x) = g(2^j x) = \sum_{k\in\mathbb{Z}} c_k \phi(2^j x - k),$$

and hence $\{\phi(2^j x - k)\}_{k\in\mathbb{Z}}$ or, equivalently, $\{\phi_{jk}\}_{k\in\mathbb{Z}}$, form an orthonormal basis of V_j. In particular, since $\phi(x) \in V_0$ we have

$$\phi(x) = \sqrt{2} \sum_{k\in\mathbb{Z}} h_k \phi(2x - k). \tag{126}$$

The coefficients in this expansion satisfy

$$\sum_k h_k = \sqrt{2}, \qquad \sum_k h_k h_{k-2l} = \delta_{0l}.$$

We then define another function (the mother wavelet)

$$\psi(x) = \sqrt{2} \sum g_k \phi(2x - k)$$

and require that $\psi(x-m)$ is orthogonal to $\phi(x)$ for all integers m, and that $\{\psi(x-m) : m \in \mathbb{Z}\}$ is an orthonormal set. These conditions impose constraints on the coefficients $\{g_k\}$

$$\sum_k g_k h_{k+2m} = 0 \;\forall m \in \mathbb{Z}, \qquad \sum_k g_k g_{k-2l} = \delta_{0l}$$

which can be satisfied by the choice $g_k = (-1)^{1-k} h_{1-k}$. It is clear that the space of functions spanned by $\{\psi(x-m) : m \in \mathbb{Z}\}$, which we denote W_0, is orthogonal to that spanned by $\{\phi(x-m) : m \in \mathbb{Z}\}$, which is V_0. The direct sum $W_0 \oplus V_0$ can be seen to coincide with V_1 (we will not prove this here, but roughly speaking V_1 is twice the size of V_0 so it makes sense that adding two orthogonal spaces of the same size as V_0 together can generate V_1).

We can continue this procedure to larger j. For each $j \geq 0$, we define the *"difference" space* W_j: $V_{j+1} = V_j \oplus W_j$, for which an orthonormal basis is given by $\{\psi_{jk}(x) : k \in \mathbb{Z}\}$. We see that $L^2(\mathbb{R}) = V_0 \oplus W_1 \oplus W_2 \oplus \ldots \oplus W_j \oplus \ldots$, and the set $\{\phi(x),\ \psi_{jk}(x) : j = 0, 1, 2, .., \ k \in \mathbb{Z}\}$ forms an orthonormal basis of $L^2(\mathbb{R})$.

10.5.5 Filter characterisation of the wavelet transform

We now prove some of the results used to describe the MRA above.

Proposition 10.4. *1.* $\sum_{k\in\mathbb{Z}} h_k = \sqrt{2}$, $\quad \sum_{k\in\mathbb{Z}} g_k = 0$

2. $\sum_{k\in\mathbb{Z}} h_k^2 = 1$, $\quad \sum_{k\in\mathbb{Z}} g_k^2 = 1$

3. For all $\ell \neq 0$, $\quad \sum_{k\in\mathbb{Z}} h_k h_{k-2\ell} = 0$, $\quad \sum_{k\in\mathbb{Z}} g_k g_{k-2\ell} = 0$

4. For all $\ell \in \mathbb{Z}$, $\quad \sum_{k\in\mathbb{Z}} g_k h_{k-2\ell} = 0$.

Proof of Properties 1 and 2. 1. To prove $\sum_{k\in Z} h_k = \sqrt{2}$, we integrate the scaling equation:

$$\begin{aligned} 1 &= \int \phi(x)dx = \sum_{k\in Z} h_k \sqrt{2} \int \phi(2x-k)dx = [v = 2x-k] = \sum_{k\in Z} h_k 2^{-1/2} \int \phi(v)dv \\ &= \frac{1}{\sqrt{2}} \sum_{k\in Z} h_k \end{aligned}$$

which implies the result.

Similarly, to prove $\sum_{k\in Z} g_k = 0$, we integrate the wavelet equation:

$$\begin{aligned} 0 &= \int \psi(x)dx = \sqrt{2} \sum_{k\in Z} g_k \int \phi(2x-k)dx = [v = 2x-k] = 2^{-1/2} \sum_{k\in Z} g_k \int \phi(v)dv \\ &= 2^{-1/2} \sum_{k\in Z} g_k \end{aligned}$$

which implies that $\sum_{k\in Z} g_k = 0$.

2. To prove $\sum_{k\in Z} h_k^2 = 1$, we integrate the squared scaling equation:

$$\begin{aligned} 1 &= \int \phi(x)^2 dx = 2 \int \left[\sum_{k\in Z} h_k \phi(2x-k) \right]^2 dx = \sum_{k,m} h_k h_m \int \phi(2x-k)\phi(2x-m)d(2x) \\ &= \sum_k h_k^2 \end{aligned}$$

since $\int \phi(2x-k)\phi(2x-m)d(2x) = 1$ if $k = m$ and is 0 otherwise.

$\sum_{k\in Z} g_k^2 = 1$ is proved similarly, by integrating the squared wavelet equation. □

The two filter decompositions (for $\phi(x)$, with coefficients $\{h_k\}$ and $\psi(x)$ with coefficients $\{g_k\}$ satisfying $g_k = (-1)^k h_{1-k}$) have other properties which we will use later to show that a finite dimensional version of wavelet decomposition, a discrete wavelet transform performed via the cascade algorithm, transforms iid Gaussian random variables to iid Gaussian random variables.

Example 10.11. *Determine filters g_k, h_k for the Haar wavelet transform.*

For the Haar wavelets, the scaling equation is

$$\mathbf{1}_{(0,1]}(x) = \mathbf{1}_{(0,1/2]}(x) + \mathbf{1}_{(1/2,1]}(x) = \mathbf{1}_{(0,1]}(2x) + \mathbf{1}_{(0,1]}(2x-1)$$

That is,

$$\phi(x) = \phi(2x) + \phi(2x-1) = \sqrt{2}\sum_{k\in\mathbb{Z}} h_k \phi(2x-k)$$

which implies that the only nonzero values of h_k are $h_0 = h_1 = 1/\sqrt{2}$.

The Haar wavelet function satisfies the following:

$$\psi(x) = \mathbf{1}_{(0,1/2]}(x) - \mathbf{1}_{(1/2,1]}(x) = \mathbf{1}_{(0,1]}(2x) - \mathbf{1}_{(0,1]}(2x-1) = \frac{1}{\sqrt{2}}\left(\phi(2x) - \phi(2x-1)\right)$$

which implies that $g_0 = 1/\sqrt{2}$, $g_1 = -1/\sqrt{2}$ and the remaining g_k are 0.

10.5.6 Discrete wavelet transform (DWT)

In typical realistic settings, we observe only a finite number of noisy values of the function. How can we obtain (noisy) wavelet coefficients based on this partial information?

10.5.7 Motivation

We want to discretise the wavelet transform:

$$\theta_{jk} = \int_0^1 f(x)\psi_{jk}(x)dx \approx \frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(i/n)f(i/n) = \frac{1}{\sqrt{n}}(Wf_n)_{(jk)} = \frac{w_{jk}}{\sqrt{n}} =: \tilde{\theta}_{jk},$$

where W, an $n\times n$ matrix defined by $W_{1i} = \phi(x_i)$, $W_{li} = \psi_{jk}(x_i)$ with $l = 2^j + k + 1$, is (approximately) orthonormal and f_n is a vector $f_n = (f(1/n), \ldots, f(1))$. We assume $n = 2^J$ for some integer J. The subscript (jk) in the above denotes the row, $l = 2^j + k + 1$, corresponding to a particular pair (j,k).

If the function f is bounded, the approximate wavelet coefficients $\tilde{\theta}_{jk}$ are close to the exact coefficients θ_{jk}: $|\tilde{\theta}_{jk} - \theta_{jk}| \leq C/n$. For Haar wavelets, $\theta_{jk} = \tilde{\theta}_{jk}$ since the Haar wavelets are constants on each interval $(i/n, (i+1)/n)$ for $n = 2^J$ for some integer J.

Use the linear transform defined by a matrix W as a discrete wavelet transform. There are other ways to derive the approximation, so that $|\tilde{\theta}_{jk} - \theta_{jk}| \leq C/n$ and matrix W is orthonormal ($WW^T = I$). In practice, it is done via the **cascade algorithm** which is derived from filter properties of wavelet transform. In this case, $|\tilde{\theta}_{jk} - \theta_{jk}| \leq C/n$ and the matrix W satisfies $WW^T = I$ due to the filter properties (Proposition 10.4).

Applying the discretised wavelet transform W to data yields

$$\begin{aligned} d_{jk} &= w_{jk} + \varepsilon_{jk}, \quad 0 \leq j \leq J-1,\ k = 0, \ldots, 2^j - 1, \\ c_{00} &= u_{00} + \varepsilon_0, \end{aligned}$$

where d_{jk} and c_{00} are discrete wavelet and scaling coefficients of observations (y_i), and ε_{jk} and ε_0 are discrete wavelet coefficients of the noise (ϵ_i). If $\epsilon_i \sim N(0,\sigma^2)$ independent, then $\varepsilon_{jk} \sim N(0,\sigma^2)$ and $\varepsilon_0 \sim N(0,\sigma^2)$ independently due to $WW^T = I$.

10.5.8 Cascade algorithm

The wavelet and scaling equations are the basis for the cascade algorithm that can be used to calculate the wavelet coefficients. The algorithm is very fast, taking $2n$ steps where n is the number of the observations. The algorithm is constructed by using recurrent equations for wavelet and scaling coefficients that are derived from the wavelet and the scaling equations in the following way.

Suppose we observe values of $f(x_i)$, $x_i = i/n$, $i = 1,\dots,n$. Denote the corresponding "noiseless" discrete scaling coefficients by u_{jk} and discrete wavelet coefficients by w_{jk} (recall that $\theta_{jk} \approx w_{jk}/\sqrt{n}$ and $\theta_0 \approx u_{00}/\sqrt{n}$). Then, the wavelet coefficients satisfy the following (using the wavelet equation):

$$\begin{aligned}
\theta_{jk} &= \int_0^1 f(x)\psi_{jk}(x)dx = \int_0^1 f(x)\psi(2^j x - k)2^{j/2}dx \\
&= \int_0^1 f(x)\left[\sqrt{2}\sum_{m\in Z} g_m\phi\left(2(2^j x - k) - m\right)\right]2^{j/2}dx \\
&= \int_0^1 f(x)\left[\sum_{m\in Z} g_m\phi\left(2^{j+1}x - 2k - m\right)2^{(j+1)/2}\right]dx \\
&= \sum_{m\in Z} g_m \int_0^1 f(x)\phi_{j+1,2k+m}(x)dx.
\end{aligned}$$

Here, $\int_0^1 f(x)\phi_{jk}(x)dx$ are scaling coefficients of f that are not used directly for estimation but are useful for computational purposes. For the discrete wavelet and scaling coefficients w_{jk} and u_{jk}, we can write the following recurrence relation:

$$w_{jk} = \sum_{m\in Z} g_m u_{j+1,2k+m}.$$

Using the scaling equation, we can derive a similar connection between the scaling coefficients at consecutive levels j and $j+1$:

$$u_{jk} = \sqrt{n}\int_0^1 f(x)\phi_{jk}(x) = \sum_{m\in Z} h_m u_{j+1,2k+m}.$$

These recurrence equations are used in the cascade algorithm. They also apply to noisy scaling and wavelet coefficients c_{jk} and d_{jk}.

We need to have a starting point. Assuming that $\text{supp}(\phi) = [0,1]$, like for the Haar scaling function, the scaling coefficients at level J for $k = 0,1,..,2^J - 1$ satisfy:

$$\begin{aligned}
&\int_0^1 f(x)2^{J/2}\phi(2^J x - k)dx = 2^{J/2}\int_{k/2^J}^{(k+1)/2^J} f(x)\phi(2^J x - k)dx \\
\approx\ & f((k+1)/n)\int_{k/2^J}^{(k+1)/2^J} 2^{J/2}\phi(2^J x - k)dx = [v = 2^J x - k] = f(x_{k+1})2^{-J/2}\int_0^1 \phi(v)dv \\
\approx\ & \frac{f(x_{k+1})}{\sqrt{n}}.
\end{aligned}$$

Therefore, we can set $u_{J,k} = f(x_{k+1})$, $k = 0, 1, \dots, 2^J - 1 = n - 1$. For noisy observations (Y_i), we can start with noisy discrete scaling coefficients $c_{J,k} = Y_{k+1}$.

Assumptions for the cascade algorithm.

1. Y_i are (noisy) observations of a function f at points x_i, $i = 1, .., n$
2. points (x_i) form a regular fixed design ($x_i - x_{i-1} = \frac{1}{n}$).
3. $n = 2^J$ for some integer J.

Cascade algorithm

1. Set $c_{Jk} = Y_{k+1}$ for $k = 0, 1, .., 2^J - 1$, set $j = J - 1$;
2. Set
$$c_{jk} = \sum_{m \in \mathbb{Z}} h_m c_{j+1,2k+m}, \quad d_{jk} = \sum_{m \in \mathbb{Z}} g_m c_{j+1,2k+m};$$
3. if $j = 0$ stop; else set $j := j - 1$ and repeat step 2.

Output: discrete wavelet coefficients c_{00}, d_{jk} for $0 \le j \le J - 1, \quad k = 0, \dots, 2^j - 1$.

Using the expressions for the Haar wavelet filters h_k and g_k, the recurrent step of the cascade algorithm for the Haar wavelet transform is

$$u_{jk} = \frac{1}{\sqrt{2}} \left(u_{j+1,2k} + u_{j+1,2k+1} \right), \quad w_{jk} = \frac{1}{\sqrt{2}} \left(u_{j+1,2k} - u_{j+1,2k+1} \right).$$

To reconstruct the function from the wavelet coefficients, this algorithm can be inverted.

10.5.9 Summary

- The number of data points $n = 2^J$.
- Cascade algorithm: set $c_{J0} = Y_1, \dots, c_{J,2^J-1} = Y_n$, and compute recursively
$$c_{jk} = \sum_m h_m c_{j+1,2k+m}, \quad d_{jk} = \sum_m g_m c_{j+1,2k+m}.$$
- The output of the the cascade algorithm are discrete wavelet coefficients: c_{00} & d_{jk}, $j < J$ that satisfy
$$d_{jk} \sim N(w_{jk}, \sigma^2), \qquad c_{00} \sim N(u_{00}, \sigma^2), \quad \text{independently.}$$
- To construct an estimator of f, choose estimators $\widehat{w}_{jk}$, $\widehat{u}_{00} (= c_{00})$, and hence construct the corresponding estimators
$$\widehat{\theta}_0 = \frac{\widehat{u}_{00}}{\sqrt{n}}, \qquad \widehat{\theta}_{jk} = \frac{\widehat{w}_{jk}}{\sqrt{n}}.$$
These estimators are then used to obtain an estimator of the function f:
$$\widehat{f}(x) = \widehat{\theta}_0 \phi(x) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \widehat{\theta}_{jk} \psi_{jk}(x).$$

For example, a linear projection estimator $\widehat{f_{J_0}}(x)$ for $f(x)$ can be constructed using the output of the cascade algorithm:

$$\widehat{w}_{jk} = d_{jk},\ j \leq J_0 - 1; \quad \widehat{w}_{jk} = 0,\ j \geq J_0; \quad \widehat{u}_{00} = c_{00}.$$

For Haar wavelets, the linear projection estimator $\widehat{f_{J_0}}$ coincides with the wavelet estimator based on discrete wavelet coefficients with $\widehat{w}_{jk} = d_{jk}$ for $j \leq J_0 - 1$ and $\widehat{w}_{jk} = 0$ for $j > J_0$.

10.5.10 Thresholding Estimators for threshold $\boldsymbol{\lambda}$

Hard thresholding estimator

$$\widehat{w}_{jk} = d_{jk} I\left(|d_{jk}| > \lambda\right) = \begin{cases} d_{jk}, & \text{if } |d_{jk}| > \lambda \\ 0, & \text{if } |d_{jk}| < \lambda \end{cases}$$

Soft thresholding estimator

$$\widehat{w}_{jk} = \begin{cases} d_{jk} - \lambda, & d_{jk} > \lambda \\ 0, & -\lambda \leq d_{jk} \leq \lambda \\ d_{jk} + \lambda, & d_{jk} < -\lambda \end{cases}$$

There is a default choice of threshold λ that is called the *universal threshold*:

$$\lambda = \sigma\sqrt{2\log n}.$$

In practice, the standard deviation σ is estimated as the median absolution deviation (MAD):

$$\widehat{\sigma} = 1.4826\ \mathrm{MAD}(d_{J-1,0}, \ldots, d_{J-1,2^J-1})$$

where $\mathrm{MAD}(x_1, \ldots, x_n) = \mathrm{median}(|x_i - \mathrm{median}(x_i)|)$.

10.5.11 Inference on f using wavelet estimators

10.5.12 Asymptotic confidence intervals for $f(x)$

$$Y_i = f(x_i) + \varepsilon_i, \quad x_i = \frac{i}{n} \quad \varepsilon_i \sim N(0, \sigma^2)$$

To construct an asymptotic confidence interval for $f(x)$, we use the linear estimator

$$\widehat{f_{J_0}(x)} = \widehat{\theta}_0\phi(x) + \sum_{j=0}^{J_0-1}\sum_{k=0}^{2^{j_0}-1} \widehat{\theta}_{jk}\psi_{jk}(x),$$

where

$$\widehat{\theta}_0 = \frac{1}{\sqrt{n}}\widehat{u}_{00}, \quad \widehat{u}_{00} = c_{00} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i\phi(x_i)$$

$$\widehat{\theta}_{jk} = \frac{1}{\sqrt{n}}\widehat{w}_{jk}, \quad \widehat{w}_{jk} = d_{jk} = \frac{1}{n}\sum_{i=1}^{n} Y_i\psi_{jk}(x_i)$$

Recall that this estimator is linear:

$$\Rightarrow \widehat{f_{J_0}(x)} = \sum_{i=1}^{n} w_i(x) Y_i, \quad w_i(x) = \frac{1}{n}\phi(x_i)\phi(x) + \frac{1}{n}\sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1} \psi_{jk}(x_i)\psi_{jk}(x),$$

therefore, given independent observations of $Y_i \sim N(f(x_i), \sigma^2)$ for $i = 1, \ldots, n$,

$$\widehat{f_{J_0}(x)} \sim N\left(f(x), \sigma^2 \sum_{i=1}^{n} w_i^2(x)\right) \quad \text{for large } n.$$

For Haar wavelets, we derived that $\sum_{i=1}^{n} w_i^2(x) = 2^{J_0}/n$.

Therefore, an asymptotic $(1-\alpha)100\%$ confidence interval for $f(x)$ based on the Haar wavelets projection estimator $\widehat{f_{J_0}}(x)$, assuming that J_0 is large enough so that the bias is much smaller than the variance, is

$$\widehat{f_{J_0}}(x) \pm z_{\alpha/2}\frac{2^{J_0/2}\sigma}{\sqrt{n}}.$$

Note that if J_0 is too large, then the confidence interval is large. Therefore, there is a tradeoff between bias and variance that results in "optimal" choice of J_0. This is discussed by considering the MISE of $\widehat{f_{J_0}}(x)$.

10.5.13 Hypothesis testing

Local support of the wavelet basis is useful when it is of interest to test whether a function is a constant on a certain subinterval of $[0, 1]$. We want to test the hypothesis

$$H_0 : f(x) = \textit{constant} \text{ on } (a, b)$$

using Haar wavelets.

Due to the support of ψ_{jk} being $[k/2^j, (k+1)/2^j]$, for $(a, b) = (m2^{-\ell}, (m+1)2^{-\ell})$ for some positive integers m and ℓ this hypothesis is equivalent to the following hypothesis about the Haar wavelet coefficients of function f:

$$H_0 : \theta_{jk} = 0 \text{ for } (j, k) \text{ such that } a < \frac{k + 1/2}{2^j} < b$$

that is, the change point of ψ_{jk} is inside (a, b). The equivalent null hypothesis can also be written as

$$H_0 : w_{jk} = 0 \text{ for } (j, k) \text{ such that } a < \frac{k + 1/2}{2^j} < b$$

since $(\theta_{jk} = w_{jk}/\sqrt{n})$ for Haar wavelets.

Test this hypothesis using observed discrete wavelet coefficients $d_{jk} \sim N(w_{jk}, \sigma^2)$, $j = 0, \ldots, J-1$, $k = 0, \ldots, 2^j - 1$, independently.

Given only $n = 2^J$ observations, we can test this hypothesis only using the wavelet coefficients with $j < J$:

$$H_0 : w_{jk} = 0 \text{ for } (j, k) \text{ such that } a < \frac{k + 1/2}{2^j} < b \,\&\, j < J.$$

Test statistic:

$$T = \sigma^{-2} \sum_{j,k:\, a<\frac{k+1/2}{2^j}<b,\, j<J} d_{jk}^2$$

which has a χ_m^2 distribution under the null hypothesis where m is the number of coefficients tested to be zero, that is, $m = \mathrm{Card}\{(j,k) : a < \frac{k+1/2}{2^j} < b,\ 0 \le j < J,\ 0 \le k \le 2^j - 1\}$.

Example 10.12. *Data:* $\mathbf{y} = (-1.0, -0.2, 0.8, 0.6, 0.0, -0.4, -0.3, -0.5)$, $x_i = i/8$, $i = 1, .., 8$, $n = 8$. *The data follows the nonparametric regression model with* $\sigma = 0.2$.

1. Test $H_0 : f(x) = const$ *on* $(1/4, 1/2)$.

Corresponding hypothesis for the wavelet coefficients is $H_0 : w_{jk} = 0$ *for* (j,k) *that satisfy* $1/4 < \frac{k+1/2}{2^j} < 1/2$, $j < J - 1 = 2$ *then* $(2^j/4 - 1/2) < k < 2^j/2 - 1/2$

Since $n = 8 = 2^3$, *we have* $J = 3$ *and hence we consider* $0 \le j \le 2$:

$j = 2$: $1/2 < k < 3/2$, *i.e.* $k = 1$ *and hence* $(j,k) = (2,1)$ *satisfies the condition*

$j = 1$: $0 < k < 1/2$ *no integer in the interval, so none*

$j = 0$: $-1/4 < k < 0$ *none.*

Therefore, the equivalent hypothesis is $H_0 : w_{21} = 0$. *Since the corresponding noisy discrete Haar wavelet coefficient* $d_{21} \sim N(w_{21}, \sigma^2)$, *under the null hypothesis* $T = d_{21}^2/\sigma^2 \sim \chi_1^2$, *therefore we reject* H_0 *at a 5% significance level if* $T = d_{21}^2/\sigma^2 > \chi_1^2(5\%) = 3.841$. *Since for this data* $d_{21} = 0.1414$ *and hence* $T = d_{21}^2/\sigma^2 = 0.5 < 3.841$, *there is not sufficient data to reject the null hypothesis at a 5% significance level.*

2. Now test $H_0 :\ f(x) = const$ *on* $(1/2, 1)$.

The corresponding hypothesis for the wavelet coefficients is $H_0 : w_{jk} = 0$ *for* (j,k) *s.t.* $1/2 < \frac{k+1/2}{2^j} < 1$, *that is, for* (j,k) *such that*

$\Leftrightarrow 2^j/2 - 1/2 < k < 2^j - 1/2$.

$j \le J - 1 = 2$. *Check this condition for each* $0 \le j \le 2$:

$j = 2$: $3/2 < k < 7/2$, *that is,* $k = 2, 3$

$j = 1$: $1/2 < k < 3/2$, *that is,* $k = 1$

$j = 0$: $0 < k < 1/2$ *none*

Therefore, the equivalent hypothesis is

$$H_0 : w_{11} = w_{22} = w_{23} = 0.$$

The test statistic is $T = (d_{11}^2 + d_{22}^2 + d_{23}^2)/\sigma^2 \sim \chi_3^2$ *under* H_0. *That is, we reject the null hypothesis at a 5% significance level if* $T > \chi_3^2(5\%) = 7.815$. *For this data,* $T = (0.2^2 + 0.2828427^2 + 0.1414214^2)/0.04 = 3.5 < 7.815$, *therefore there is not sufficient data to reject the null hypothesis at a 5% significance level.*

Remark 10.2. *For an arbitrary interval* (a,b) *(that is, not of the form* $(m2^{-\ell}, (m+1)2^{-\ell})$*), the equivalent null hypothesis in terms of Haar wavelet coefficients is*

$$H_0 : w_{jk} = 0 \text{ for } (j,k) \text{ such that } \{a < \frac{k}{2^j} < b \text{ or } a < \frac{k+1/2}{2^j} < b \text{ or } a < \frac{k+1}{2^j} < b\},$$

for $j = 0, 1, \ldots, J-1$ *and* $k = 0, 1, \ldots, 2^j - 1$. *That is, in the more general case we need to check if any of the three points where the Haar wavelet* ψ_{jk} *jumps between different constant values is inside the interval* (a,b).

For an interval of the type $(m2^{-\ell}, (m+1)2^{-\ell})$ *it is not necessary to check the end point since they are either at the same place with regard to* (a,b) *(that is, inside or outside) as the mid point* $(k+1/2)2^{-j}$ *or on the boundary of the interval.*

10.5.14 MISE (mean integrated square error) of wavelet estimators

Suppose a function f has the following wavelet decomposition:

$$f(x) = \theta_0 \phi(x) + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \theta_{jk} \psi_{jk}(x),$$

and consider a wavelet estimator

$$\hat{f}(x) = \hat{\theta}_0 \phi(x) + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \hat{\theta}_{jk} \psi_{jk}(x).$$

Lemma 10.2. *(Parseval identity). For a function f and its wavelet estimator $\hat{f}(x)$,*

$$||f - \hat{f}||_2^2 = (\theta_0 - \hat{\theta}_0)^2 + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} (\hat{\theta}_{jk} - \theta_{jk})^2.$$

This is due to the wavelet basis being orthonormal.

Consider the following estimator of the wavelet coefficients for $j = 0, .., J_0 - 1$ for some J_0:

$$\hat{\theta}_{jk} = \frac{1}{n} \sum_{i=1}^{n} \psi_{jk}(x_i) Y_i,$$

and $\hat{\theta}_{jk} = 0$ for $j \geq J_0$. The estimator of the scaling coefficient is $\hat{\theta}_0 = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) Y_i$. Sometimes we refer to θ_0 as $\theta_{-1,0}$, and to $\phi(x)$ as $\psi_{-1,0}(x)$.

The corresponding wavelet estimator is

$$\hat{f}_{J_0}(x) = \sum_{j \leq J_0 - 1} \sum_{k} \hat{\theta}_{jk} \psi_{jk}(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i \sum_{j \leq J_0 - 1} \sum_{k} \psi_{jk}(x_i) \psi_{jk}(x).$$

This wavelet estimator

$$\hat{f}_{J_0}(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i \sum_{j \leq J_0 - 1} \sum_{k} \psi_{jk}(x_i) \psi_{jk}(x)$$

is linear since it can be written as

$$\hat{f}_{J_0}(x) = \sum_{i=1}^{n} Y_i W_i(x),$$

with $W_i(x) = \frac{1}{n} \sum_{j \leq J_0 - 1, k} \psi_{jk}(x_i) \psi_{jk}(x)$, i.e., that is independent of the Y_i's.

By Lemma 10.2,

$$\mathbb{E}||f - \hat{f}||_2^2 = \mathbb{E}(\theta_0 - \hat{\theta}_0)^2 + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \mathbb{E}(\hat{\theta}_{jk} - \theta_{jk})^2,$$

hence it is sufficient to find MSE of $\hat{\theta}_{jk}$, $\mathbb{E}(\hat{\theta}_{jk} - \theta_{jk})^2$.

We know that

$$\mathbb{E}(\hat{\theta}_{jk}-\theta_{jk})^2 = \mathrm{Var}(\hat{\theta}_{jk}) + \left[\mathrm{bias}(\hat{\theta}_{\mathrm{jk}})\right]^2 .$$

Therefore, we need to find the variance and the bias of $\hat{\theta}_{jk}$.

Variance

For $j \le J_0 - 1$,

$$\begin{aligned}\mathrm{Var}\left(\hat{\theta}_{jk}\right) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(x_i)Y_i\right)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\psi_{jk}^2(x_i)\mathrm{Var}\left(Y_i\right) = \frac{\sigma^2}{n}\frac{1}{n}\sum_{i=1}^{n}\psi_{jk}^2(x_i)\\ &= \frac{\sigma^2}{n}(1+o(1)),\end{aligned}$$

due to the independence of the Y_i's and $\frac{1}{n}\sum_{i=1}^{n}\psi_{jk}^2(x_i) \approx \int_0^1 \psi_{jk}^2(x)dx = 1$.

Bias

For $j \le J_0 - 1$, the bias is

$$\mathbb{E}\left(\hat{\theta}_{jk}-\theta_{jk}\right) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)\psi_{jk}(x_i) - \int_0^1 f(x)\psi_{jk}(x)dx.$$

Assume that $f \in H^{\beta}(M_f)$ and is bounded, i.e. $|f(x)| \le C_f$ for all $x \in [0,1]$. We assume that the wavelet function ψ is such that $|\psi(x)-\psi(y)| \le M_{\psi}|x-y|$ for all $x, y \in [0,1]$, and it is bounded: $|\psi(x)| \le C_{\psi}$ for all $x \in [0,1]$ (and that the same conditions hold for the scaling function ϕ). We also assume that $\mathrm{supp}(\psi) \subseteq [0,1]$ and $\mathrm{supp}(\phi) \subseteq [0,1]$.

Under these assumptions with $\beta \in (0,1]$, the absolute value of the bias is bounded by

$$\begin{aligned}|\mathbb{E}\left(\hat{\theta}_{jk}-\theta_{jk}\right)| &\le \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} |f(x)\psi_{jk}(x) - f(x_i)\psi_{jk}(x_i)|\,dx\\ &\le \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} \left[|f(x)\psi_{jk}(x) - f(x)\psi_{jk}(x_i)| + |f(x)\psi_{jk}(x_i) - f(x_i)\psi_{jk}(x_i)|\right]dx\\ &\le \max_x |f(x)| 2^{j/2}\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left|\psi(2^jx-k)-\psi(2^jx_i-k)\right|dx\\ &\quad +\sum_{i=1}^{n}|\psi_{jk}(x_i)|\int_{x_{i-1}}^{x_i}|f(x)-f(x_i)|\,dx.\end{aligned}$$

Considering the first term on the right hand side, we have

$$\begin{aligned}\int_{x_{i-1}}^{x_i}\left|\psi(2^jx-k)-\psi(2^jx_i-k)\right|dx &\le M_{\psi}\int_{x_{i-1}}^{x_i}|2^jx-k-(2^jx_i-k)|dx\\ &\le 0.5M_{\psi}2^jn^{-2}.\end{aligned}$$

The intersection of the interval of integration $[(i-1)/n, i/n]$ and the support of ψ_{jk}

$$\mathrm{supp}(\psi_{jk}) = [k2^{-j},(k+1)2^{-j}] = [k2^{J-j}/n,(k+1)2^{J-j}/n]$$

is nonempty (and consists of more than a single point) iff $k2^{J-j} < i-1 < (k+1)2^{J-j}$ or $k2^{J-j} < i < (k+1)2^{J-j}$, i.e. $k2^{J-j}+1 \leq i <\leq (k+1)2^{J-j}$. There are 2^{J-j} of such i. Thus,

$$\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}|\psi_{jk}(x)-\psi_{jk}(x_i)|dx \leq 0.5M_\psi 2^j n^{-2}2^{J-j} = 0.5M_\psi n^{-2}2^J = 0.5M_\psi n^{-1},$$

using $n = 2^J$ and hence

$$\max_x |f(x)|2^{j/2}\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left|\psi(2^jx-k)-\psi(2^jx_i-k)\right|dx \leq 0.5C_fM_\psi 2^{j/2}n^{-1}.$$

For the second term, we have

$$\int_{x_{i-1}}^{x_i}|f(x)-f(x_i)|\,dx \quad\leq\quad M_f\int_{x_{i-1}}^{x_i}|x-x_i|^\beta \leq \frac{M_f}{(\beta+1)n^{\beta+1}},$$

and using the restriction to the support of ψ_{jk}

$$|\psi_{jk}(x_i)| \leq 2^{j/2}C_\psi \mathbf{1}(k2^{J-j}+1 < i < (k+1)2^{J-j}),$$

$$\Rightarrow \sum_{i=1}^{n}|\psi_{jk}(x_i)| \leq 2^{j/2}C_\psi\sum_{i=1}^{n}\mathbf{1}(k2^{J-j}+1 \leq i \leq (k+1)2^{J-j}) \leq 2^{J-j/2}C_\psi \leq C_\psi n2^{-j/2}.$$

Thus,

$$|\mathbb{E}\hat{\theta}_{jk}-\theta_{jk}| \quad\leq\quad 0.5C_fM_\psi 2^{j/2}n^{-1}+\frac{M_fC_\psi}{(\beta+1)}2^{-j/2}n^{-\beta}$$

again using $n = 2^J$ and $j < J$.

MSE $(\hat{\theta}_{jk})$ **for** $j \geq J_0$

For $j \geq J_0$, $\hat{\theta}_{jk} = 0$, and therefore the MSE $(\hat{\theta}_{jk}) = \mathbb{E}(\hat{\theta}_{jk}-\theta_{jk})^2 = \theta_{jk}^2$.

For $f \in H^\beta(M_f)$, $|\theta_{jk}| \leq M_f2^{-j(\beta+1/2)}$ for all j, k.

Now we summarise the properties of **bias and variance of** $\hat{\theta}_{jk}$ that we have derived.

Lemma 10.3. *Assume that*

- $f \in H^\beta(M_f)$, $\beta \in (0,1)$, *and* $|f(x)| \leq C_f$ *for all* $x \in [0,1]$;
- ψ *is such that supp*$(\psi) \subseteq [0,1]$, $|\psi(x)-\psi(y)| \leq M_\psi|x-y|$ *for all* $x, y \in [0,1]$, *and it is bounded:* $|\psi(x)| \leq C_\psi$ *for all* $x \in [0,1]$ *(and that the same conditions hold for the scaling function* ϕ*).*

Then, for $\hat{\theta}_{jk} = \frac{1}{n}\sum_{i=1}^n \psi_{jk}(x_i)Y_i$,

$$\begin{aligned} Var\left(\hat{\theta}_{jk}\right) &= \frac{\sigma^2}{n}(1+o(1)) \quad \textit{as } n \to \infty, \\ |\mathrm{bias}(\hat{\theta}_{\mathrm{jk}})| &\leq c_12^{j/2}n^{-1}+c_22^{-j/2}n^{-\beta}, \end{aligned}$$

where $c_1 = 0.5C_fM_\psi$ *and* $c_2 = \frac{M_fC_\psi}{(\beta+1)}$.

MISE of $\hat{f}_{J_0}(x)$
Under the assumptions of Lemma 10.3, the MISE of the linear wavelet estimator is

$$\begin{aligned}
\mathbb{E}||f - \hat{f}_{J_0}||_2^2 &= \mathbb{E}(\theta_0 - \hat{\theta}_0)^2 + \sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1}\mathbb{E}(\hat{\theta}_{jk} - \theta_{jk})^2 + \sum_{j=J_0}^{\infty}\sum_{k=0}^{2^j-1}\theta_{jk}^2 \\
&\leq 2^{J_0}\frac{\sigma^2}{n}(1+o(1)) + 2c_1^2 n^{-2}[1 + \sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1} 2^j] \\
&\quad +2c_2^2 n^{-2\beta}[1 + \sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1} 2^{-j}] + M_f^2 \sum_{j=J_0}^{\infty}\sum_{k=0}^{2^j-1} 2^{-j(2\beta+1)} \\
&= 2^{J_0}\frac{\sigma^2}{n}(1+o(1)) + 2c_1^2 n^{-2}(2^{2J_0}+2)/3 + 2c_2^2 n^{-2\beta}(J_0+1) + M_f^2\frac{2^{-2\beta J_0}}{1-2^{-2\beta}} \\
&\leq \sigma^2\frac{N}{n}(1+o(1)) + \tilde{c}_1 n^{-2}N^2 + \tilde{c}_2 n^{-2\beta}\log n + \tilde{c}_3 N^{-2\beta} + \tilde{c}_4 n^{-2}
\end{aligned}$$

where $N = 2^{J_0} < 2^J = n$ and $\tilde{c}_1 = 2c_1^2/3$, $\tilde{c}_2 = 2c_2^2$, $\tilde{c}_3 = M_f^2(1-2^{-2\beta})^{-1}$ and $\tilde{c}_4 = 4c_1^2/3$.

For the estimator to be consistent, we need the MISE to tend to 0 as $n \to \infty$, therefore we need $N/n \to 0$ and $N \to \infty$ as $n \to \infty$. In this case, the second term is much smaller than the first one, and $\log N < \log n$. Therefore, to find the optimal N (and hence the optimal J_0) that minimises the upper bound on the MISE, we can consider just 2 remaining terms:

$$MISE(\hat{f}_{J_0}) \leq \sigma^2\frac{N}{n}(1+o(1)) + \tilde{c}_3 N^{-2\beta}(1+o(1))$$

This expression is minimised when $N = cn^{1/(2\beta+1)}$, that is, when $2^{J_0} = c2^{J/(2\beta+1)}$ which implies that $J_0 = \frac{J}{2\beta+1}(1+o(1))$ as $n \to \infty$ (and hence as $J \to \infty$).

Therefore, the linear wavelet estimator with $J_0 = \frac{J}{2\beta+1}$ has MISE bounded by

$$\text{MISE}(\hat{f}_{J_0}) \leq Cn^{-2\beta/(2\beta+1)}$$

that is, it achieves the global minimax rate of convergence, and it has the same rate of convergence as the kernel estimator with the optimal bandwidth.

Note that this estimator is non-adaptive, that is, we need to know β, the smoothness of the unknown function, to estimate f well. The wavelet thresholding estimator with the threshold $(1+d)\sigma\sqrt{2\log n}$ for any $d \in (0,1)$ (that is, slightly larger than the universal threshold) achieves the optimal rate of convergence (up to a factor of $\log n$) **adaptively**, that is, without using the smoothness of f.

11 Gaussian and Dirichlet Processes

We encountered stochastic processes when we discussed noise in gravitational wave detectors and then again in the discussion of Time Series. Another application of stochastic processes is to generate probability distributions, as the relative frequencies of different outcomes of the stochastic process over long time intervals. We will be concerned with two particular types of stochastic process.

- **Gaussian processes:** These are infinite dimensional generalisations of the Normal distribution and realisations of these are random fields.
- **Dirichlet processes:** These are infinite dimensional generalisations of the Dirichlet distribution, and realisations of these are probability distributions.

11.1 Gaussian processes

A multivariate Gaussian distribution returns values of a finite set of random variables. A natural extension is to regard the set of random variables as the values of some random field at certain points. To generate the full random field we need an infinite dimensional Gaussian distribution, which is a Gaussian process. Formally we denote a random field, $y(\mathbf{x})$, generated by a Gaussian process via

$$y(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

where $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ are the mean and covariance function of the Gaussian process. For simplicity of notation we assume that the random field is single valued at each point, but the extension to multivariate outputs is straightforward.

Formally, a GP is an infinite collection of variables, any finite subset of which are distributed as a multivariate Gaussian. For a set of parameter points $\{\mathbf{x}_i\}$, including, but not limited to, the training set $\mathcal{D}$,

$$[y(\mathbf{x}_i)] \sim N(\boldsymbol{m}, \boldsymbol{K}) \,, \tag{127}$$

where the mean vector and covariance matrix of this Gaussian distribution are fixed by the corresponding functions of the GP,

$$[\boldsymbol{m}]_i = m(\mathbf{x}_i) \,, \quad [\boldsymbol{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \,, \tag{128}$$

with probability density function

$$P\left(\{y(\mathbf{x}_i\}\right) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{K}|}} \exp\left(-\frac{1}{2}\sum_{i,j}(y(\mathbf{x}_i) - m(\mathbf{x}_i))\left[\boldsymbol{K}^{-1}\right]_{ij}(y(\mathbf{x}_j) - m(\mathbf{x}_j))\right) \,. \tag{129}$$

Gaussian processes are often used for interpolation. In that context, the training set $\mathcal{D}$ represents the set of known values of the field, e.g., the results of computational simulations at certain choices of input parameters, which we denote by $\tilde{y}(\mathbf{x}_i)$. The Gaussian process is constrained by this training set and then used to predict the value of the field at new points in the parameter space, with associated uncertainties. If the values of the field at the training points are not known perfectly, but have uncertainties $\epsilon_i \sim N(0, \sigma_i^2)$, the expression above takes the same form but with the replacement

$$[\boldsymbol{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) + \sigma_i^2 \delta_{ij}.$$

Even with perfect simulations it can be advantageous to include a small error term, as this helps with inversion of the covariance matrix.

The mean and variance of the GP determine how the function is interpolated across the parameter space. It is common in regression to set the mean of the Gaussian process to zero, but specifying the covariance function is central to GP regression as it encodes our prior expectations about the properties of the function being interpolated. Possibly the simplest and most widely used choice for the covariance function is the squared exponential (SE)

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left[-\frac{1}{2} g_{ab} (\mathbf{x}_i - \mathbf{x}_j)^a (\mathbf{x}_i - \mathbf{x}_j)^b\right] , \tag{130}$$

which defines a stationary, smooth GP. In Eq. (130), a scale σ_f and a (constant) metric g_{ab} for defining a modulus in parameter space have been defined. These are called hyperparameters and we denote them as $\vec{\theta} = \{\sigma_f, g_{ab}\}$, with Greek indices μ, ν, ... to label the components of this vector.

The probability in Eq. (129) is referred to as the hyperlikelihood, or alternatively the evidence for the training set; it is the probability that that particular realisation of waveform differences was obtained from a GP with a zero mean and specified covariance function. The hyperlikelihood depends only on the hyperparameters and the quantities in the training set, so we denote it as $Z(\vec{\theta}|\mathcal{D})$. The log hyperlikelihood is

$$\begin{aligned} \ln Z(\vec{\theta}|\mathcal{D}) \quad = \quad & -\frac{N}{2}\ln(2\pi) \\ & -\frac{1}{2}\sum_{i,j} (y(\mathbf{x}_i) - m(\mathbf{x}_i)) \left[k(\mathbf{x}_i, \mathbf{x}_j)\right]^{-1} (y(\mathbf{x}_j) - m(\mathbf{x}_j)) \\ & -\frac{1}{2}\ln\left|\det\left[k(\mathbf{x}_i, \mathbf{x}_j)\right]\right| . \end{aligned} \tag{131}$$

The values of the hyperparameters can be fixed to their optimum values $\vec{\theta}_{\mathrm{op}}$, defined as those which maximise the hyperlikelihood:

$$\left.\frac{\partial Z(\vec{\theta}|\mathcal{D})}{\partial \theta^\mu}\right|_{\vec{\theta}=\vec{\theta}_{\mathrm{op}}} = 0 . \tag{132}$$

An alternative approach is to consider the hyperparameters as nuisance parameters in addition to the source parameters $\mathbf{x}$, and marginalise over them while sampling an expanded likelihood,

$$\Lambda_{\mathrm{expanded}}(\mathbf{x}, \vec{\theta}|\mathcal{D}) \quad \propto \mathcal{L}(\mathbf{x}|\vec{\theta}, \mathcal{D}) Z(\vec{\theta}|\mathcal{D}). \tag{133}$$

The disadvantage of this approach is that the hyperlikelihood is expensive to compute and the inclusion of extra nuisance parameters slows down any application of the GP. In contrast, maximising the likelihood is a convenient heuristic which is widely used in other contexts and allows all the additional computation to be done offline.

Having fixed the properties of the covariance function by examining the training set, we can now move on to using the GP as a predictive tool. The defining property of the GP is that any finite collection of variables drawn from it is distributed as a multivariate Gaussian in the manner of Eq. (129). Therefore, the set of variables formed by the training set plus the field at a set of extra parameter points $\{y(\mathbf{z}_j)\}$ is distributed as

$$\begin{bmatrix} y(\mathbf{x}_i) \\ y(\mathbf{z}_j) \end{bmatrix} \sim N\left(\boldsymbol{m}, \boldsymbol{\Sigma}\right) \,, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{K} & \boldsymbol{K}_* \\ \boldsymbol{K}_*^{\mathrm{T}} & \boldsymbol{K}_{**} \end{pmatrix} \,, \tag{134}$$

where $\boldsymbol{K}$ is defined in Eq. (128) and the matrices $\boldsymbol{K}_*$ and $\boldsymbol{K}_{**}$ are defined as

$$[\boldsymbol{K}_*]_{ij} = k(\mathbf{x}_i, \mathbf{z}_j) \,, \quad [\boldsymbol{K}_{**}]_{ij} = k(\mathbf{z}_i, \mathbf{z}_j) \,. \tag{135}$$

The conditional distribution of the unknown field values at the new points, given the observed values in $\mathcal{D}$, can now be found and is given by

$$p(\{y(\mathbf{z}_i)\}) \propto \exp\left[-\frac{1}{2}\sum_{j,k}(y(\mathbf{z}_j) - \mu_j)\Sigma^{-1}_{jk}(y(\mathbf{x}_k) - \mu_k)\right] \tag{136}$$

where the GPR mean and its associated error are given by

$$\mu_i \quad = \quad m(\mathbf{z}_i) + \sum_{j,k} [\boldsymbol{K}_*]_{ji} \left[\boldsymbol{K}^{-1}\right]_{jk} (\tilde{y}(\mathbf{x}_k) - m(\mathbf{x}_k)) \,, \tag{137}$$

$$\Sigma_{ij} \quad = \quad [\boldsymbol{K}_{**}]_{ij} - \sum_{k,l} [\boldsymbol{K}_*]_{ki} \left[\boldsymbol{K}^{-1}\right]_{kl} [\boldsymbol{K}_*]_{lj} \,. \tag{138}$$

11.2 The covariance function

The properties of the covariance function play an important role in determining the nature of the Gaussian process and its behaviour when used for regression. The only necessary requirements we have of a covariance function are that it is a positive definite; i.e. for any choice of points $\{\mathbf{x}_i\}$ the covariance matrix $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ is positive definite. The covariance function (and the corresponding GP) is said to be stationary if the covariance is a function only of $\vec{\tau} = \mathbf{x}_1 - \mathbf{x}_2$, furthermore it is said to be isotropic if it is a function only of $\tau \equiv |\vec{\tau}| = |\mathbf{x}_1 - \mathbf{x}_2|$.[3] Isotropy of a GP implies stationarity, but the converse is not true.

An example of how the properties of the covariance function relate to the properties of the GP, and hence the properties of the resulting interpolant, is given by considering the mean-square (MS) continuity and differentiability of GPs. It can be shown that the first ζ MS derivatives of a GP are MS continuous (the GP is said to be ζ-times MS differentiable) if and only if the first 2ζ derivatives of the covariance function are continuous at the diagonal point $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_*$. For a stationary GP this condition reduces to checking the 2ζ derivatives of $k(\vec{\tau})$ at $\vec{\tau} = \vec{0}$, and for an isotropic GP checking the 2ζ derivatives of $k(\tau)$ at $\tau = 0$.

It is the smoothness properties of the covariance function at the origin that determine the differentiability of the GP. In the following subsections, we consider two aspects that enter the definition of the covariance function:

1. specifying the distance metric in parameter space g_{ab};

2. specifying the functional form of the covariance with distance $k(\tau)$,

These cannot be completely separated; there exists an arbitrary scaling, α of the distance $\tau \to \alpha\tau$ which can be absorbed into the definition of the covariance, $k(\tau) \to k(\tau/\alpha)$. However, provided the steps are tackled in order, there is no ambiguity.

[3] We have yet to define a metric on parameter space with which to take the norm of this vector (see Sec. 11.2.2), but all that is required here is that a suitably smooth metric exists.

11.2.1 The metric g_{ab}

One simple way to define a distance τ between two points in parameter space, and the way used in the SE covariance function in Eq. (130), is to define $\tau^2 = g_{ab}(\mathbf{x}_1 - \mathbf{x}_2)^a(\mathbf{x}_1 - \mathbf{x}_2)^b$, where g_{ab} are constant hyperparameters. This distance is obviously invariant under a simultaneous translation of $\mathbf{x}_1 \to \mathbf{x}_1 + \boldsymbol{\Delta}$ and $\mathbf{x}_2 \to \mathbf{x}_2 + \boldsymbol{\Delta}$; therefore, this defines a stationary GP. For a D-dimensional parameter space, this involves specifying $D(D+1)/2$ hyperparameters g_{ab}.

More complicated distance metrics (with a larger number of hyperparameters) are possible if the condition of stationarity is relaxed, i.e. $g_{ab} \to g_{ab}(\mathbf{x})$. Given a family of stationary covariance functions, a non-stationary generalisation can be constructed. A stationary covariance function can be considered as a kernel function centred at $\mathbf{x}_1$; $k(\mathbf{x}_1, \mathbf{x}_2) \equiv k_{\mathbf{x}_1}(\mathbf{x}_2)$. Allowing a different kernel function to be defined at each point $\mathbf{x}_1$, a new, non-stationary covariance function is $k(\mathbf{x}_1, \mathbf{x}_2) = \int \mathrm{d}\vec{u}\, k_{\vec{u}}(\vec{\lambda_1}) k_{\vec{u}}(\mathbf{x}_2)$.[4] Applying this procedure to a D-dimensional SE function generates a non-stationary analogue

$$\begin{aligned} k(\mathbf{x}_i, \mathbf{x}_j) &= \sigma_f \left|\mathcal{G}^i\right|^{1/4} \left|\mathcal{G}^j\right|^{1/4} \left|\frac{\mathcal{G}^i + \mathcal{G}^j}{2}\right|^{-1/2} \\ &\quad \times \exp\left(-\frac{1}{2} Q_{ij}\right) , \end{aligned} \tag{139}$$

where

$$Q_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^a (\mathbf{x}_i - \mathbf{x}_j)^b \left(\frac{\mathcal{G}^i_{ab} + \mathcal{G}^j_{ab}}{2}\right)^{-1} , \tag{140}$$

and $\mathcal{G}^i_{ab} = \mathrm{inv}[g_{ab}(\mathbf{x}_i)]$ is the inverse of the parameter-space metric at position $\mathbf{x}_i$. Provided that the metric $g_{ab}(\mathbf{x})$ is smoothly parameterised this non-stationary SE function retains the smoothness properties discussed earlier.

The generalisation in Eq. (139) involves the inclusion of a large set of additional hyperparameters to characterise how the metric changes over parameter space; for example one possible parameterisation would be the Taylor series

$$g_{ab}(\mathbf{x}) = g_{ab}(\mathbf{x}_0) + (\mathbf{x}^c - \mathbf{x}^c_0) \left.\frac{\partial g_{ab}(\mathbf{x})}{\partial \lambda^c}\right|_{\mathbf{x}=\mathbf{x}_0} + \ldots \tag{141}$$

with the hyperparameters $g_{ab}(\mathbf{x}_0)$, $\partial g_{ab}(\mathbf{x})/\partial \lambda^c$, and so on. The inclusion of even a single extra hyperparameter can incur a significant Occam penalty which pushes the training set to favour a simpler choice of covariance function. For this reason most applications use stationary GPs.

An alternative to considering non-stationary metrics is instead to try and find new coordinates $\tilde{\lambda} \equiv \tilde{\lambda}(\mathbf{x})$ such that the metric in these coordinates becomes (approximately) stationary. Such transformations are very problem specific and finding them typically requires expert knowledge of the context of the application.

[4] To see that k is a valid covariance function consider an arbitrary series of points $\{\mathbf{x}_i\}$, and the sum over training set points $I = \sum_{i,j} a_i a_j k(\mathbf{x}_i, \mathbf{x}_j)$; for k to be a valid covariance it is both necessary and sufficient that $I \geq 0$. Using the definition of k gives $I = \int \mathrm{d}\vec{u} \sum_{i,j} a_i a_j k_{\vec{u}}(\vec{\lambda_i}) k_{\vec{u}}(\mathbf{x}_j) = \int \mathrm{d}\vec{u} \, (\sum_i a_i k_{\vec{u}}(\vec{\lambda_i}))^2 \geq 0$.

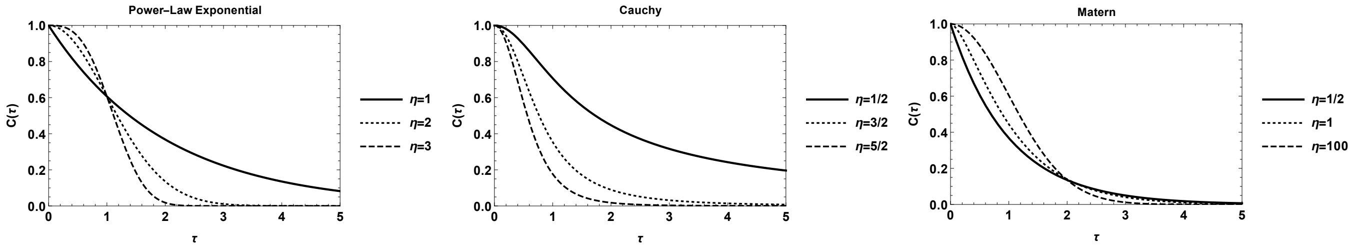


Figure 47: Plots of the different generalisations of the SE covariance function discussed in Sec. 11.2.2. The left-hand panel shows the PLE function, the centre panel shows the Cauchy function, and the right-hand panel shows the Matérn function; in all cases the value of σ_f was fixed to unity. In each panel the effect of varying the additional hyperparameter is shown by the three curves. For the PLE covariance the case $\eta = 2$ recovers the SE covariance, while for the Cauchy and Matérn covariances the case $\eta \to \infty$ recovers the SE covariance.

11.2.2 The functional form of $k(\tau)$

The second stage of specifying the covariance function involves choosing the function of distance $k(\tau)$. In general whether a particular function $k(\tau)$ is positive definite (and hence is a valid covariance function) depends on the dimensionality D of the underlying space (i.e. $\mathbf{x} \in \mathbb{R}^D$); however, all the functions considered in this section are valid for all D. Several choices for $k(\tau)$ are particularly common in the literature, these include the SE covariance function (which has already been introduced), given by

$$k_{\rm SE}(\tau) = \sigma_f^2 \exp\left(-\frac{1}{2}\tau^2\right) . \tag{142}$$

The power-law exponential (PLE) covariance function, given by

$$k_{\rm PLE}(\tau) = \sigma_f^2 \exp\left(-\frac{1}{2}\tau^\eta\right) , \tag{143}$$

where $0 < \eta \leq 2$. The PLE reduces to the SE in the case $\eta = 2$. The Cauchy function, given by

$$k_{\rm Cauchy}(\tau) = \frac{\sigma_f^2}{(1+\tau^2/2\eta)^\eta} , \tag{144}$$

where $\eta > 0$. This recovers the SE function in the limit $\eta \to \infty$. And finally, the Matérn covariance function, given by

$$k_{\rm Mat}(\tau) = \frac{\sigma_f^2 2^{1-\eta}}{\Gamma(\eta)} \left(\sqrt{2\eta}\,\tau\right)^\eta K_\eta\left(\sqrt{2\eta}\,\tau\right) , \tag{145}$$

where $\eta > 1/2$, and K_η is the modified Bessel function of the second kind [?]. In the limit $\eta \to \infty$, the Matérn covariance function also tends to the SE.

Fig. 47 shows the functional forms of the covariance functions. They have similar shapes; they return a finite covariance at zero distance which decreases monotonically, and tends to zero as the distance becomes large. In the case of regression this indicates that the values of the field at two nearby points in parameter space are closely related, whereas the values at two well separated points are nearly independent. The PLE, Cauchy and Matérn function can all be viewed as attempts to generalise the SE with the inclusion of one extra hyperparameter η, to allow for more flexible GP modelling. All three alternative functions are able to recover the SE in some limiting case, but the Matérn is the most flexible of the three. This can be seen from the discussion of the MS differentiability of GPs given in section 11.3.

The SE covariance function is infinitely differentiable at $\tau = 0$, and so the corresponding GP is infinitely MS differentiable. The PLE function is infinitely differentiable at $\tau = 0$ for the SE case when $\eta = 2$, but for all other cases it is not at all MS differentiable. In contrast, the Cauchy function is infinitely differentiable at $\tau = 0$ for all choices of the hyperparameter η. The Matérn function, by contrast, has a variable level of differentiability at $\tau = 0$, controlled via the hyperparameter η. The GP corresponding to the Matérn covariance function in Eq. (145) is ζ-times MS differentiable if and only if $\eta > \zeta$. This ability to modify the differentiability allows the same covariance function to successfully model a wide variety of data. In the process of maximising the hyperlikelihood for the training set over hyperparameter η, the GP learns the (non)smoothness properties favoured by the data, and the the GPR returns a correspondingly (non)smooth function.

11.2.3 Compact support and sparseness

All of the covariance functions considered up until this point have been strictly positive;

$$k(\tau) > 0 \quad \forall \tau \in [0, \infty) \,. \tag{146}$$

When evaluating the covariance matrix for the training set K_{ij} this leads to a matrix where all entries are positive definite; i.e. a dense matrix. When performing the GPR it is necessary to maximise the hyperlikelihood for the training set with respect to the hyperparameters. This process involves inverting the dense matrix K_{ij} at each iteration of the optimisation algorithm. Although this procedure is carried out offline, it still can become prohibitive for large training sets. A related problem, as pointed out in Sec. **??** is that for large training sets the determinant of the covariance matrix is typically small which also contributes to making the covariance matrix hard to invert.

One potential way around these issues is to consider a covariance function with compact support,

$$\begin{aligned} k(\tau) > 0 &\quad \tau \in [0, T] \,, \\ k(\tau) = 0 &\quad \forall \tau \in (T, \infty) \,, \end{aligned} \tag{147}$$

where T is some threshold distance beyond which we assume that the waveform differences become uncorrelated. This leads to a sparse, band-diagonal covariance matrix, which is much easier to invert. Care must be taken when specifying the covariance function to ensure that the function is positive definite (which is required of a GP): if the SE covariance function is truncated, then the matrix formed from the new covariance function is not guaranteed to be positive definite.

Nevertheless, it is possible to construct covariance functions which have the requisite properties and satisfy the compact support condition in Eq. (147). These are typically based on polynomials. We consider a series of polynomials, originally proposed by Wendland. These have the property that they are positive definite in $\mathbb{R}^D$ and are $2q$-time differentiable at the origin. Therefore the discrete parameter q is in some sense analogous to the η hyperparameter of the Matérn covariance function in that it controls the smoothness of the GP.

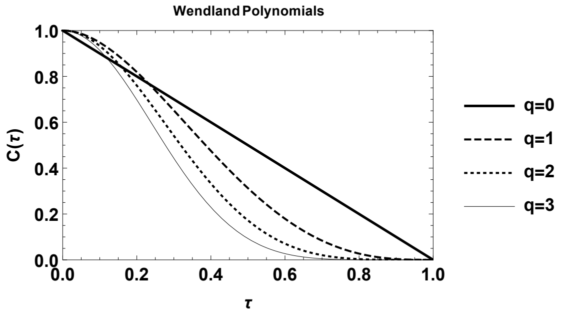


Figure 48: Plots of the first few Wendland polynomial covariance functions. All these functions have compact support, $k(\tau) = 0$ for $\tau > 1$. As the value of q increases the functions become smoother near the origin.

Defining β to be

$$\beta = \left\lfloor \frac{D}{2} \right\rfloor + q + 1 \tag{148}$$

and where $\Theta(x)$ denotes the Heaviside step function, the first few Wendland polynomials $k_{D,q}(\tau)$ are given by,

$$\begin{aligned}
k_{D,0}(\tau) &= \sigma_f^2 \Theta(1-\tau)(1-\tau)^{\beta}\,, && (149)\\
k_{D,1}(\tau) &= \sigma_f^2 \Theta(1-\tau)(1-\tau)^{\beta+1}\left[1 + (\beta+1)\,\tau\right]\,, && (150)\\
k_{D,2}(\tau) &= \frac{\sigma_f^2}{3}\Theta(1-\tau)(1-\tau)^{\beta+2}\left[\,3 + (3\beta+6)\,\tau\right.\\
&\quad \left. + \left(\beta^2+4\beta+3\right)\tau^2\right]\,, && (151)\\
k_{D,3}(\tau) &= \frac{\sigma_f^2}{15}\Theta(1-\tau)(1-\tau)^{\beta+3}\left[15 + (15\beta+45)\,\tau\right.\\
&\quad + \left(6\beta^2+36\beta+45\right)\tau^2\\
&\quad \left. + \left(\beta^3+9\beta^2+23\beta+15\right)\tau^3\right]\,. && (152)
\end{aligned}$$

The first few Wendland polynomials are plotted in Fig. 48. Other types of covariance functions with compact support have also been proposed and explored in the literature, but we do not consider them here.

11.3 Continuity and differentiability of GPs

Before moving on to some examples, we give proofs concerning the continuity and differentiability of GPs. Let $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \ldots$ be a sequence of points in parameter space which converges to a point $\mathbf{x}_*$, in the sense $\lim_{\ell\to\infty}\ |\mathbf{x}_\ell\ -\ \mathbf{x}_*| = 0$. The GP $Y(\mathbf{x})$ is said to be MS continuous at $\mathbf{x}_*$ if

$$\lim_{\ell\to\infty} \mathbb{E}\left[(Y(\mathbf{x}_\ell) - Y(\mathbf{x}_*)|Y(\mathbf{x}_\ell) - Y(\mathbf{x}_*))\right] = 0\,, \tag{153}$$

where $\mathbb{E}[\ldots]$ denotes the expectation of the enclosed quantity over realisations of the GP. MS continuity implies continuity in the mean,

$$\lim_{\ell\to\infty} \mathbb{E}\left[Y(\mathbf{x}_\ell) - Y(\mathbf{x}_*)\right] = 0\,. \tag{154}$$

This follows from considering the variance of the quantity $Y(\mathbf{x}_\ell) - Y(\mathbf{x}_*)$, and the fact that variance is non-negative. There are other notions of continuity of GPs used in the literature, but the notion of MS continuity relates most easily to the covariance.

The mean and the covariance of a GP are defined as

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[Y(\mathbf{x})] \,, \\ k(\mathbf{x}_1, \mathbf{x}_2) &= \mathbb{E}[(Y(\mathbf{x}_1) - m(\mathbf{x}_1)|Y(\mathbf{x}_2) - m(\mathbf{x}_2))] \,. \end{aligned} \tag{155}$$

Using these, Eq. (153) can be written as

$$\begin{aligned} \lim_{\ell\to\infty} \{ & k(\mathbf{x}_*, \mathbf{x}_*) - 2k(\mathbf{x}_\ell, \mathbf{x}_*) + k(\mathbf{x}_\ell, \mathbf{x}_\ell) \\ & + (m(\mathbf{x}_*) - m(\mathbf{x}_\ell)|m(\mathbf{x}_*) - m(\mathbf{x}_\ell))\} = 0 \,, \end{aligned} \tag{156}$$

and using the continuity of the mean in Eq. (154) gives

$$\lim_{\ell\to\infty} [k(\mathbf{x}_*, \mathbf{x}_*) - 2k(\mathbf{x}_\ell, \mathbf{x}_*) + k(\mathbf{x}_\ell, \mathbf{x}_\ell)] = 0 \,. \tag{157}$$

This condition is satisfied if the covariance function is continuous at the point $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_*$. Therefore, we arrive at the result that if the covariance function is continuous in the usual sense at some point $\mathbf{x}_*$, then the corresponding GP is MS continuous at this point.[5] In the special case of stationary covariance this reduces to checking continuity of $k(\vec{\tau})$ at $\vec{\tau} = 0$, and in the special case of isotropic covariance, continuity of $k(\tau)$ at $\tau = 0$.

We now move on from continuity to consider differentiability. In the spirit of Eq. (153), the notion of taking the MS derivative of a GP is defined as

$$\frac{\partial Y(\mathbf{x})}{\partial \mathbf{x}^a} = \underset{\epsilon\to 0}{\text{l.i.m}}\, X_a(\mathbf{x}, \epsilon) \,, \tag{158}$$

where l.i.m is read limit in MS and

$$X_a(\mathbf{x}, \epsilon) = \frac{Y(\mathbf{x} + \epsilon\, \hat{e}_a) - Y(\mathbf{x})}{\epsilon} \tag{159}$$

with parameter-space unit vector $\hat{e}_a$. This definition can be extended to higher-order derivatives in the obvious way.

The MS derivative of a GP is also a GP; this follows simply from the fact that the sum of Gaussians is also distributed as a Gaussian. The covariance of $X_a(\mathbf{x}, \epsilon)$ is given by

$$\begin{aligned} K_\epsilon(\mathbf{x}_1, \mathbf{x}_2) &= \mathbb{E}\,[(X_a(\mathbf{x}_1, \epsilon) - \Xi(\mathbf{x}_1, \epsilon)| \\ &\qquad X_a(\mathbf{x}_2, \epsilon) - \Xi(\mathbf{x}_2, \epsilon))] \end{aligned} \tag{160}$$

where $\Xi_a(\mathbf{x}, \epsilon) = \mathbb{E}[X_a(\mathbf{x}, \epsilon)]$. It then follows that

$$\begin{aligned} K_\epsilon(\mathbf{x}_1, \mathbf{x}_2) &= \frac{k(\mathbf{x}_1 + \epsilon, \mathbf{x}_2 + \epsilon) - k(\mathbf{x}_1, \mathbf{x}_2 + \epsilon)}{\epsilon^2} \\ &\quad + \frac{k(\mathbf{x}_1 + \epsilon, \mathbf{x}_2) - k(\mathbf{x}_1, \mathbf{x}_2)}{\epsilon^2} \,. \end{aligned} \tag{161}$$

[5] A GP is continuous in MS if and only if the covariance function is continuous, although this is not proved here.

Substituting this into Eq. (158), the limit in MS becomes a normal limit, and the result is obtained that the MS derivative of a MS continuous GP with covariance function $k(\mathbf{x}_1, \mathbf{x}_2)$ is a GP with covariance function $\partial^2 k(\mathbf{x}_1, \mathbf{x}_2)/\partial \mathbf{x}_1^a \partial \mathbf{x}_2^a$. In general the covariance function of the ζ-times MS differentiated GP

$$\frac{\partial^\zeta Y(\mathbf{x})}{\partial \mathbf{x}^{a_1} \partial \mathbf{x}^{a_2} \ldots \partial \mathbf{x}^{a_\zeta}}, \tag{162}$$

is given by the 2ζ-times differentiated covariance function

$$\frac{\partial^{2\zeta} k(\mathbf{x}_1, \mathbf{x}_2)}{\partial \mathbf{x}_1^{a_1} \partial \mathbf{x}_2^{a_1} \partial \mathbf{x}_1^{a_2} \partial \mathbf{x}_2^{a_2} \ldots \partial \mathbf{x}_1^{a_\zeta} \partial \mathbf{x}_2^{a_\zeta}}. \tag{163}$$

From the above results relating the MS continuity of GPs to the continuity of the covariance function at $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_*$, it follows that the ζ-times MS derivative of the GP is MS continuous (the GP is said to be ζ-times MS differentiable) if the 2ζ-times derivative of the covariance function is continuous at $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_*$. So it is the smoothness properties of the covariance function along the diagonal points that determine the differentiability of the GP.[6]

11.4 Example applications of Gaussian processes

Example: interpolation of a quadratic We consider first a toy problem in which we generate noisy measurements, $\{y_i\}$, at 200 points, $\{x_i\}$, randomly chosen in the interval $[0, 1]$ according to

$$y_i = -2 - 3x_i + 5x_i^2 + \epsilon_i, \qquad \epsilon_i \sim N(0, 0.15^2).$$

We then fit a Gaussian process to a training set comprising a subset of these points. We use a squared exponential covariance function and optimize the hyperparameters over the training set. The results of this procedure are shown in Figure 49. Results are represented by the expectation value and 1σ uncertainty computed from the fitted Gaussian process as a function of x. We see that the Gaussian process is well able to recover the true function, even with as a few as ten training points. This is a particularly simple function and if we knew that the relationship was quadratic there would be no need to use a Gaussian process to fit the data. In Figure 50 we show the result of fitting a quadratic model to the same data. As expected, the fit is slightly better, but not hugely so. The advantage of the Gaussian process approach is that you do not need to know the form of the model in advance, and avoid the problem of model mis-specification. In Figure **??** we show the result of fitting a linear model to the same data. We see that we end up with a very precise, but wrong, representation of the curve. Gaussian process regression models have greater flexibility and should always converge to the true underlying function in the limit that the number of observations tends to infinity.

Example: waveform model errors We will now consider a few examples from the gravitational wave literature. There are many of these that have all appeared since ∼ 2015, so we cannot describe them all but we will mention a few different examples. The first application of Gaussian processes in a gravitational wave context was to characterise uncertainties coming from waveform model errors (Moore & Gair (2014)). A Gaussian process was used to model the error in a particular waveform model family over parameter

[6] It can be further shown that if a covariance function $k(\mathbf{x}_1, \mathbf{x}_2)$ is continuous at every diagonal point $\mathbf{x}_1 = \mathbf{x}_2$ then it is everywhere continuous.

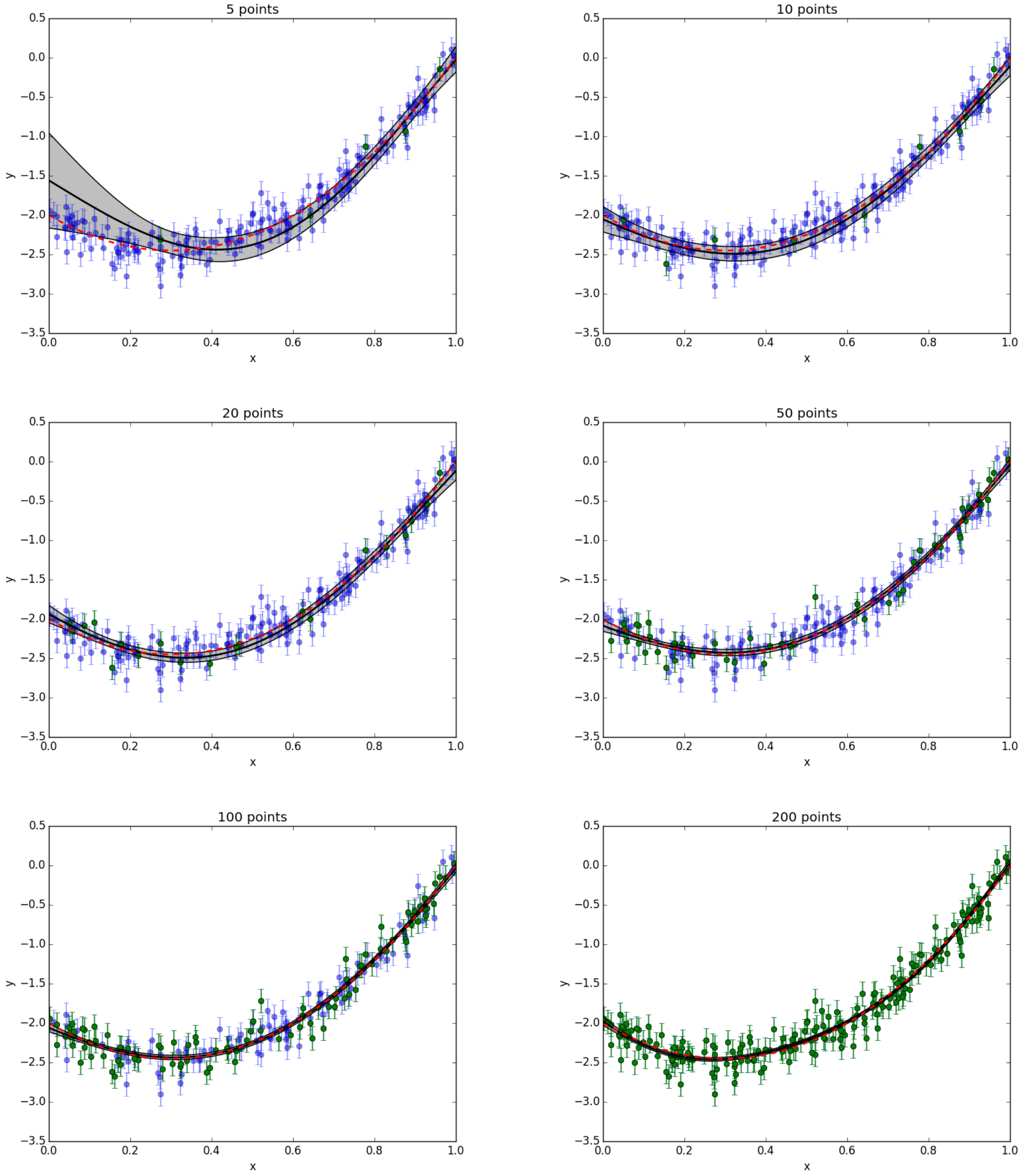


Figure 49: Gaussian process fit to noisy measurements of a quadratic, for different sizes of training set, as stated in the title of each panel.

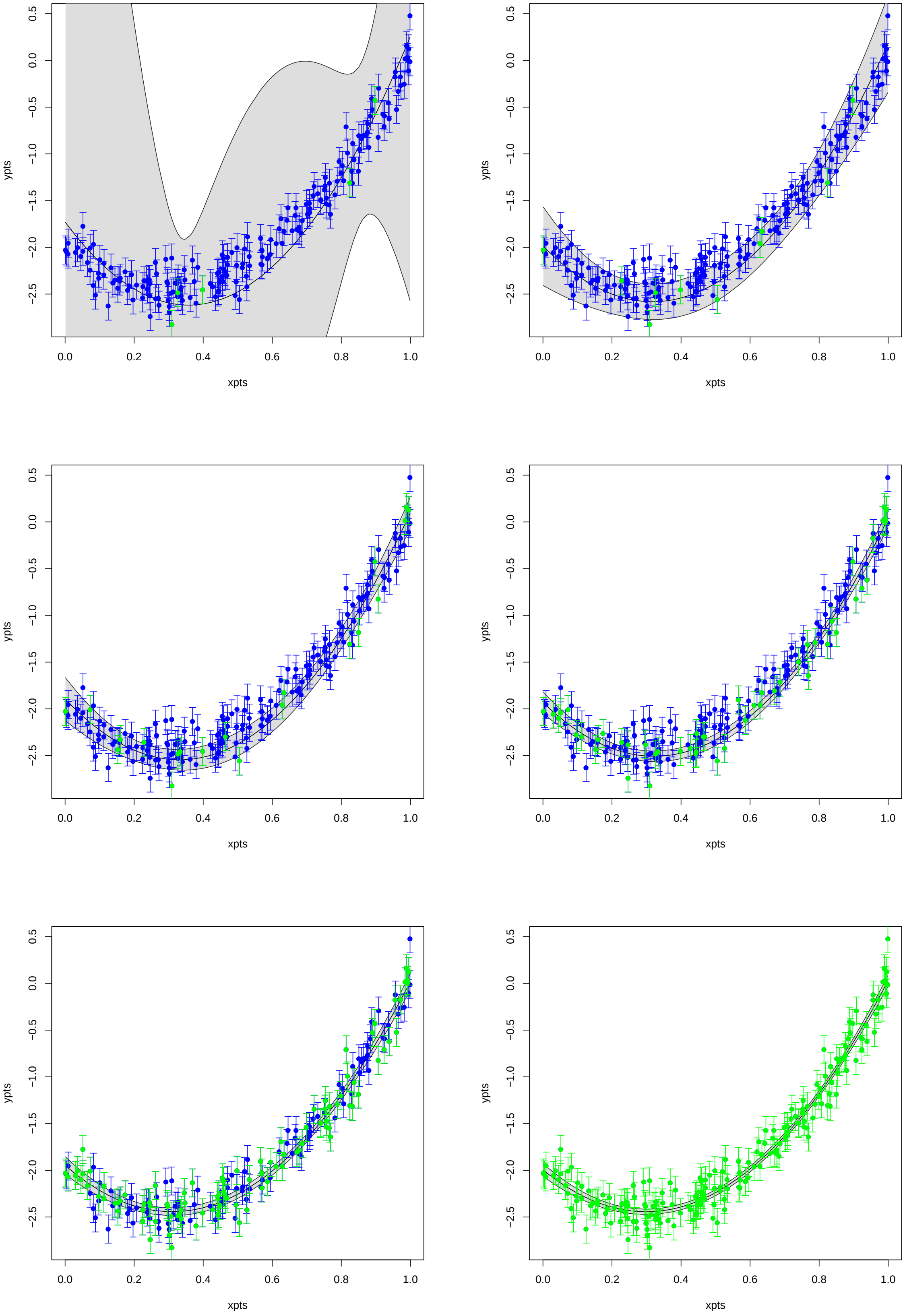


Figure 50: As Figure 49, but now fitting a quadratic linear model to the same data.

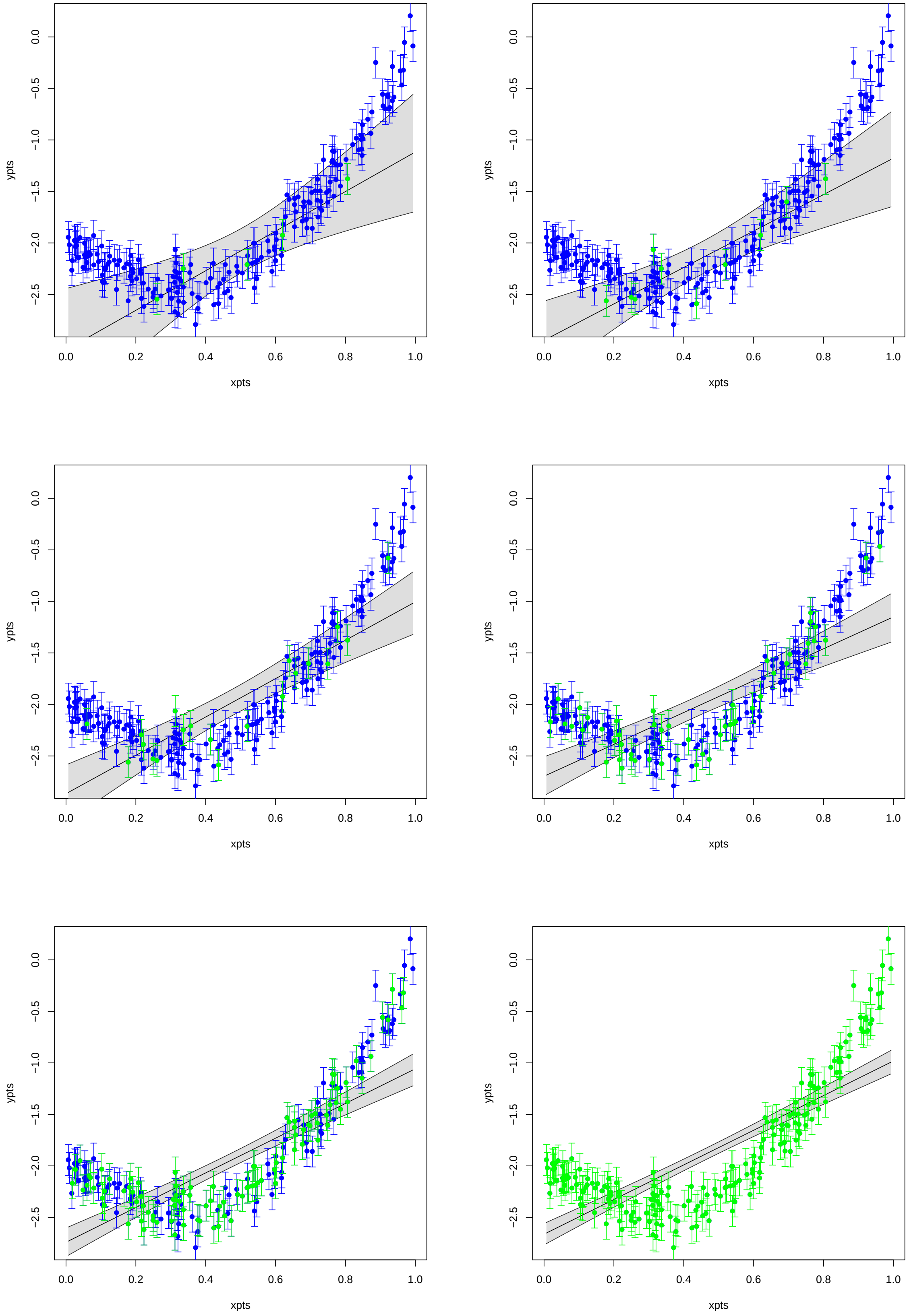


Figure 51: As Figure 49, but now fitting a linear model to the same data.

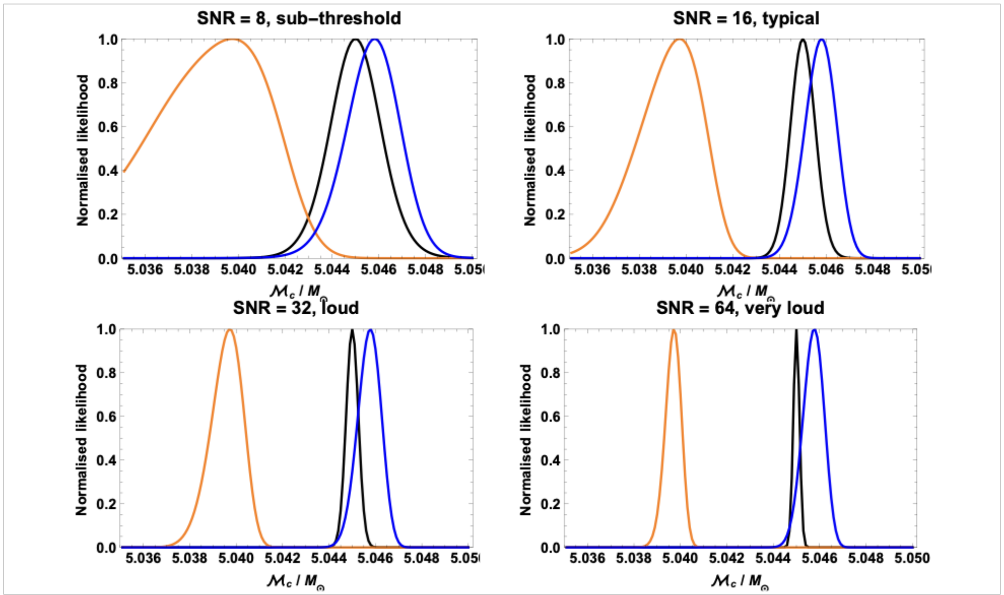


Figure 52: Comparison between uncorrected, corrected and "true" likelihood for inference with waveform models that include model error. The corrected likelihood uses a Gaussian process to model the waveform error and then marginalises this out of the likelihood. Reproduced from Moore et al. (2015).

space. Using a training set based on model errors estimated as the difference between two different approximate waveforms, a Gaussian process model for the waveform error was produced. As this distribution is Gaussian and so is the normal gravitational wave likelihood, the waveform error can then be marginalised out of the likelihood to give an alternative **marginalised likelihood** for use in parameter estimation. This marginalised likelihood took the form

$$\mathcal{L}(\vec{\lambda}) \quad \propto \quad \frac{1}{\sqrt{1+\sigma^2(\vec{\lambda})}} \exp\left(-\frac{1}{2} \frac{\left\| s - H(\vec{\lambda}) + \mu(\vec{\lambda}) \right\|^2}{1+\sigma^2(\vec{\lambda})} \right). \tag{164}$$

In this $\vec{\lambda}$ is the vector of parameters characterising the gravitational wave signal, the quantity $\mu(\vec{\lambda})$ is the Gaussian process estimate for the model error, and shifts the distribution to eliminate the error, and $\sigma^2(\vec{\lambda})$ is the variance in the Gaussian process, which widens the posterior to account for the uncertainty in the model error. Use of this marginalised likelihood corrects for biases in parameter estimation, as illustrated in Figure 52.

Example: waveform interpolation In Williams et al. (2020), Gaussian processes were used to directly model the gravitational waveform, rather than its error. A set of numerical relativity waveforms were used to create a training set to which a Gaussian process model was fitted. In Figure 53 we show some random draws from the GP model at a certain point in parameter space and compare these to two different waveform approximants evaluated at the same point. We see that the GP uncertainty band includes all of the different approximants and so automatically factors in waveform uncertainty.

Example: population inference In Taylor & Gerosa (2018), a Gaussian process was used as a means to interpolate the output of binary population synthesis code over the

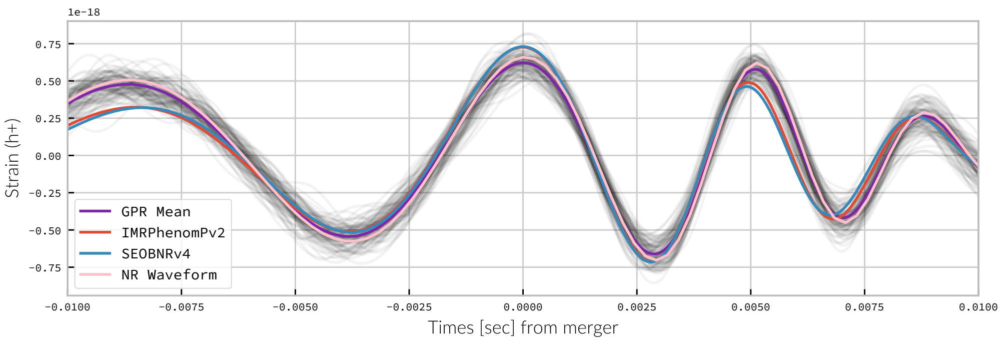


Figure 53: Comparison of several approximate waveform models to random draws from a Gaussian process interpolant trained on numerical relativity simulations. Reproduced from Williams et al. (2020).

space of physical parameters that characterise them. The resulting model, continuous over parameter space, was then used to infer properties of the underlying astrophysical population based on a set of observed compact binary inspirals. Figure **??** shows simulated inferred posteriors on the population parameters that were produced in this way.

Example: equation of state uncertainties Landry & Essick (2019) and Essick, Landry & Holz (2019) used a Gaussian process to model the equation of state of a neutron star, $p(\rho)$. The hyperparameters of the Gaussian process were constrained using a training set including numerical equation of state simulations. The resulting model generates random equations of state which can be used to marginalise equation of state uncertainties out of inference on gravitational wave signals from binary neutron stars. Figure 55 shows a set of random draws of the equation of state from the Gaussian process.

11.5 Dirichlet Processes

Recall that a Dirichlet distribution generates a set of K random values, $\{x_i\}$, constrained to take values with $0 \leq x_i \leq 1$ for all i and $\sum x_i = 1$. The distribution depends on a vector of parameters $\vec{\alpha} = (\alpha_1, \ldots, \alpha_K)$ and has pdf

$$p(\vec{x}) = \frac{1}{B(\vec{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \qquad B(\vec{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{j=1}^{K} \alpha_j\right)}.$$

A realisation of a Dirichlet distribution is a probability mass function for a discrete distribution with K possible outcomes. A **Dirichlet process** generalises the Dirichlet distribution to infinite dimensions and a realisation of a Dirichlet process is a continuous probability distribution. A Dirichlet process is characterised by a **base distribution**, P, and a **concentration parameter**, a. The base distribution is a probability measure on a set S. The process X is a Dirichlet process, denoted $X \sim$DP(P, a) if for any measurable finite partition of the set S, $\{B_i\}_{i=1}^{n}$, the probability distribution on this partition generated by X is

$$(X(B_1), X(B_2), \ldots, X(B_n)) \sim \mathrm{Dir}(aP(B_1), aP(B_2), \ldots, aP(B_n)). \qquad (165)$$

In the limit $a \to 0$, the Dirichlet pdf, which is proportional to $x_i^{\alpha_i - 1}$, places a logarithmically increasing weight towards the lower boundary of the variable range. Draws from this

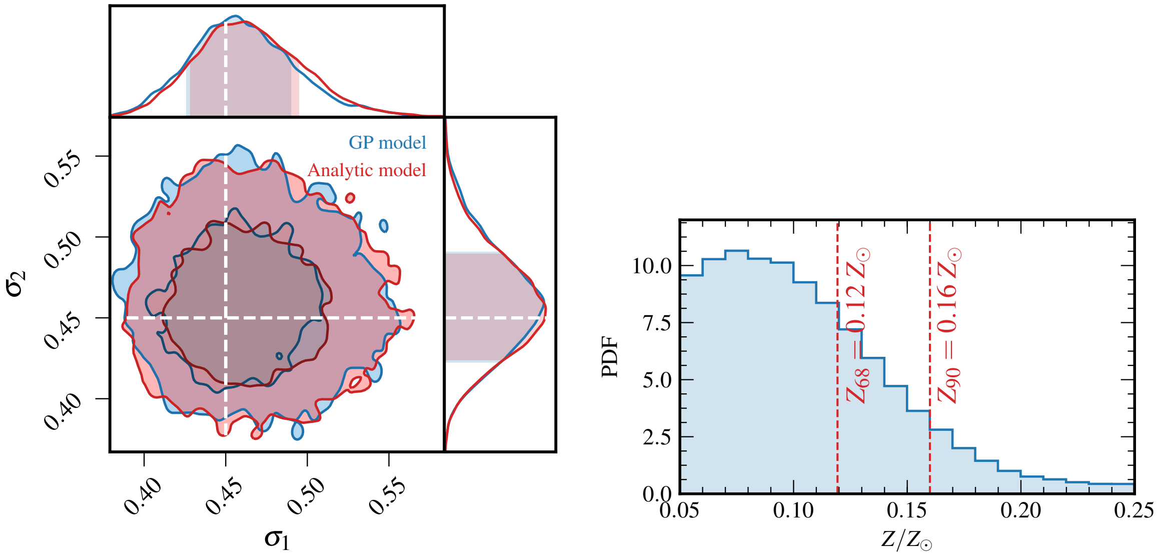


Figure 54: Posteriors on physical parameters of the astrophysical source population inferred form simulated observations of binaries. Inference relied on a Gaussian process model that interpolated the output of the population synthesis codes over the astrophysical parameter space. Reproduced from Taylor & Gerosa (2018).

distribution will therefore be singletons, with all x_i's bar one equal to zero. For small a the Dirichlet distribution will therefore tend to be discretized, with probability concentrated at a small number of locations.

In the limit $a \to \infty$, the distribution becomes more and more concentrated at its mode, which is at $x_i = P(B_i)$. Every realisation of $\mathrm{Dir}(aP(B_1), aP(B_2), \ldots, aP(B_n))$ therefore returns $(P(B_1), \ldots, P(B_n))$ and every realisation of the Dirichlet process thus gives the base distribution.

These limits show that the Dirichlet process generates discretized representations of the base distribution, with the level of discretization decreasing as $a \to \infty$. To illustrate this, we show in Figure 56 and 57 some realisations of a Dirichlet process, for a fixed base distribution, $P = N(0, 1)$, and various choices of a. In each figure, we represent the realisation of the Dirichlet process by a set of 1000 random draws from the realised probability distribution. It is clear that for small a, only a small number of values are returned, showing high discretisation, but as a increases the number of distinct values is increasing and the distribution becomes a closer and closer approximation to the base distribution.

11.5.1 Sampling Dirichlet processes

A realisation of a Dirichlet process is a probability distribution on S and hence infinite dimensional. Drawing such a realisation is therefore very difficult. However, in practice what we need is not the realisation of the Dirichlet process itself but a set of samples from that realised distribution, which is much easier to obtain. If the full realisation is required, this can be evaluated by looking at the distribution of a large number of samples. This is how the realisations shown in Figures 56 and 57 were produced.

There are several different algorithms for drawing samples from a random realisation of a Dirichlet process, $X \sim \mathrm{DP}(P, a)$. The **chinese restaurant process** generates a sequence

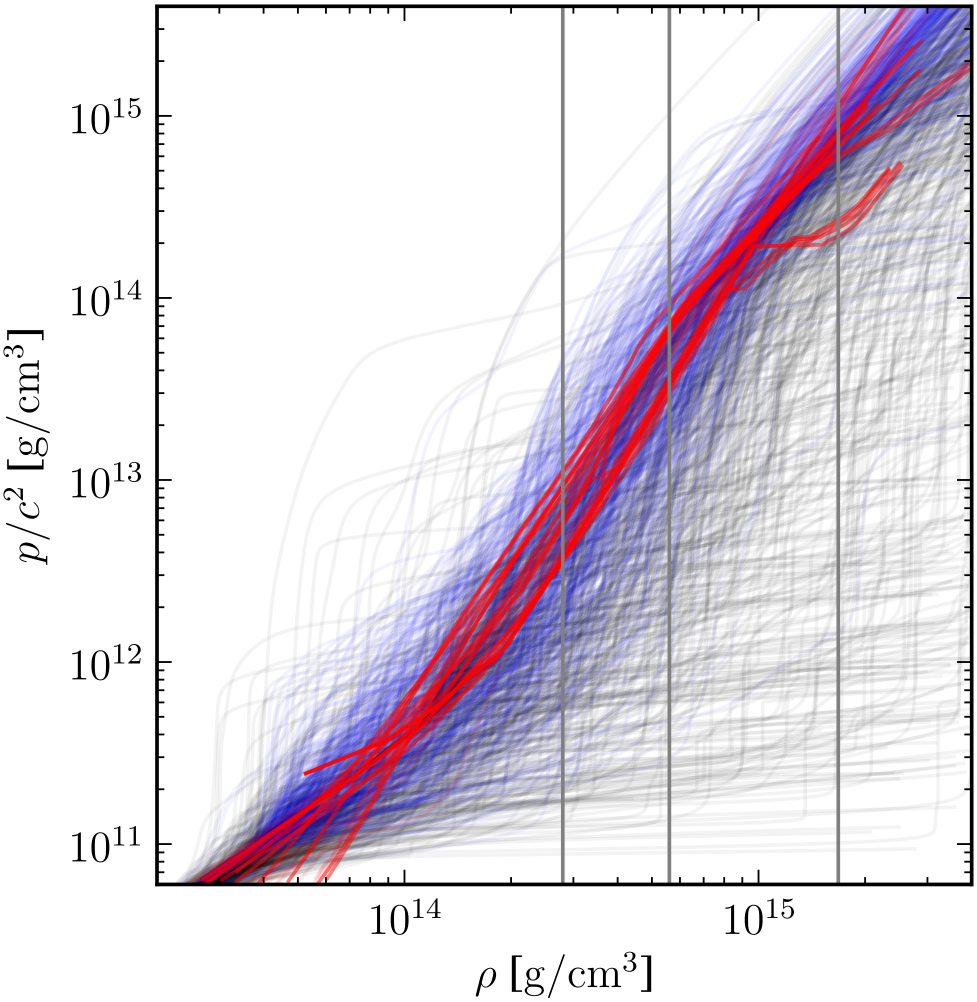


Figure 55: Random draws from a Gaussian process model of the equation of state of a neutron star. Reproduced from Essick et al. (2019).

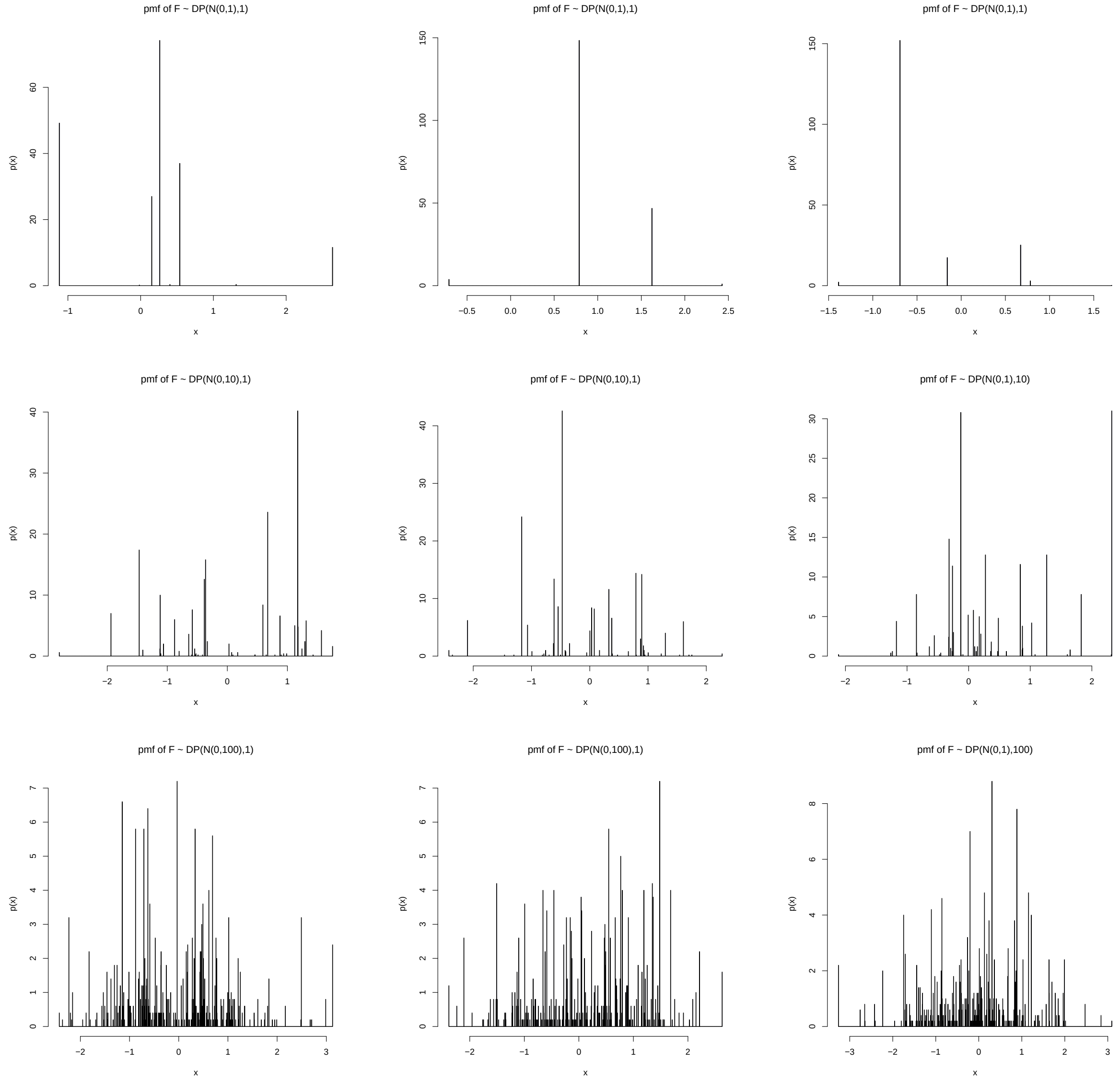


Figure 56: Sample realisations of a Dirichlet process, X $\sim$DP$(N(0,1),a)$, for $a = 1$ (top row), $a = 10$ (middle row) and $a = 100$ (bottom row). In each figure we show 1000 samples from the given realisation of the Dirichlet process. Within each row, the figures show three distinct realisations of the stated Dirichlet process.

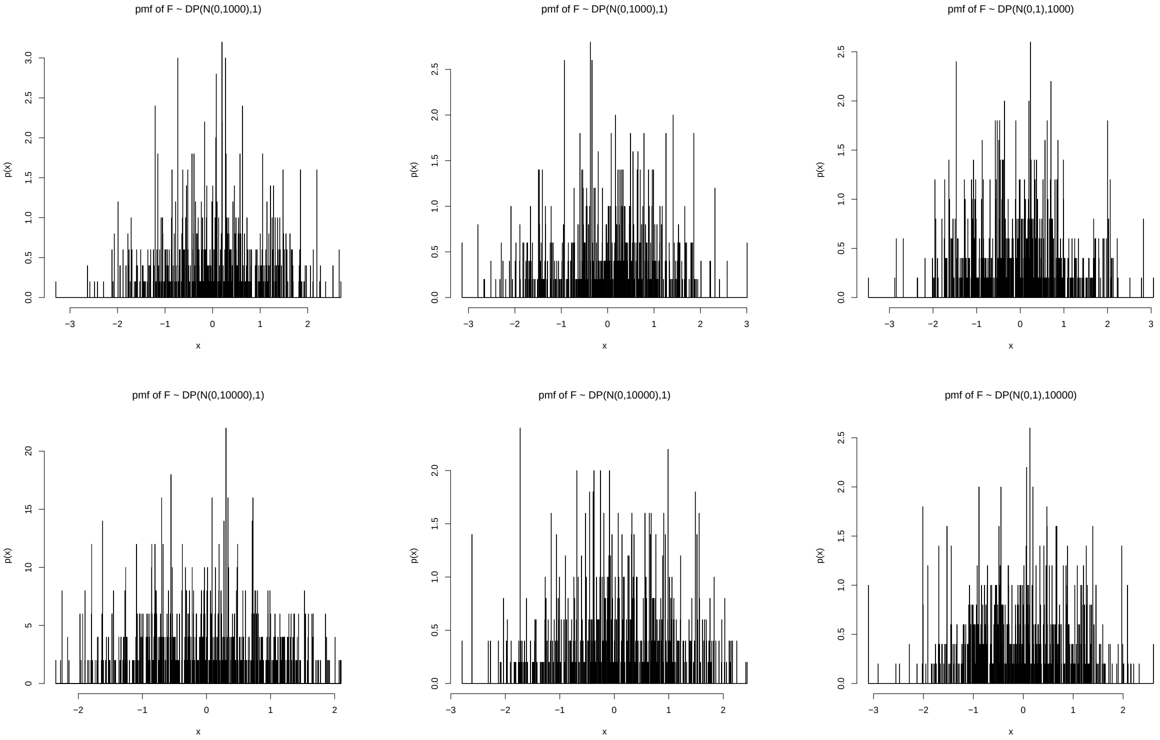


Figure 57: As in Figure 56, these figures show sample realisations of a Dirichlet process, $X \sim$DP$(N(0,1),a)$, for $a = 1000$ (top row) and $a = 10000$ (bottom row). In each figure we show 1000 samples from the given realisation of the Dirichlet process. Within each row, the figures show three distinct realisations of the stated Dirichlet process.

of samples $\{x_i\}$ for $i \geq 1$ as follows

- with probability $a/(a+i-1)$ draw x_i from P;
- with probability $n_x/(a+i-1)$ set $x_i = x$, where n_x is the number of previous observations of $x_j = x$ for $j < i$.

This procedure is called the chinese restaurant process by analogy with a restaurant with an infinite number of tables, each serving a different dish, and each with infinite seating capacity. A new diner may choose to sit at a new table, or may choose to sit at a table where people are already eating. The probability of choosing a particular table is proportional to the number of people observed already sitting at that table and enjoying the offered dish.

Closely related to this is the **Polya Urn** scheme. In that construction we start with an urn containing a black balls. At each step of the algorithm, a ball is drawn at random from the urn. If the ball is black, we generate a new color randomly, color a new ball this color and return it to the urn along with the black ball. The corresponding sample is the new color. If the ball drawn is coloured, then we take a new ball, color it the same color as the sampled ball, and return both of them to the urn. The corresponding sample is the color of the ball that was drawn. It is clear that the distribution of colors produced in this way corresponds to the samples generated form the chinese restaurant process.

A final approach to constructing a sample from a random realisation of a Dirichlet process is the **stick breaking** construction. This approach explicitly generates a discrete distribution, X, which is a realisation of the Dirichlet process. The distribution is given by

$$\begin{aligned}
&X = \left(\sum_{l=1}^{L_H} p_l \delta_{U_l}\right) + \left(1 - \sum_{l=1}^{L_H} p_l\right) \delta_{U_0} \\
&p_1 = V_1, \qquad p_l = \left(\prod_{j=1}^{l-1} (1 - V_j)\right) V_l, \;\; l \geq 2, \qquad p_0 = 1 - \sum_{l=1}^{L_H} p_l \\
&V_l \sim \text{Beta}(1, a), \;\; l = 1, \ldots, L_H, \qquad U_l \sim P, \;\; l = 0, 1, \ldots, L_H,
\end{aligned} \tag{166}$$

where we take the limit $L_H \to \infty$, but in practical applications the procedure is truncated at some finite, but sufficiently large, value.

11.5.2 Example applications

The main application of Dirichlet processes is in the field of Bayesian nonparametrics, where they are used as a prior for unknown probability distributions. We will provide two examples.

Example: B-spline regression In the nonparametric regression chapter we encountered the notion of smoothing splines for regression. In that context, the knots of the spline were fixed at the locations of the observed data points. The number of knots is therefore fixed for any given data set and grows as $n \to \infty$. The smoothing was controlled by the regularisation parameter. Another approach to nonparametric regression is to allow the number of spline points to vary and let the data choose the optimal number. Even greater flexibility comes from allowing the locations of the spline knots to vary. In Edwards & Gair (2020) they presented a Bayesian nonparametric regression algorithm that uses B-splines (an alternative basis for cubic splines than the one presented in this course), but with the number and location of the knots both allowed to vary and adapt to the data. The knot locations were

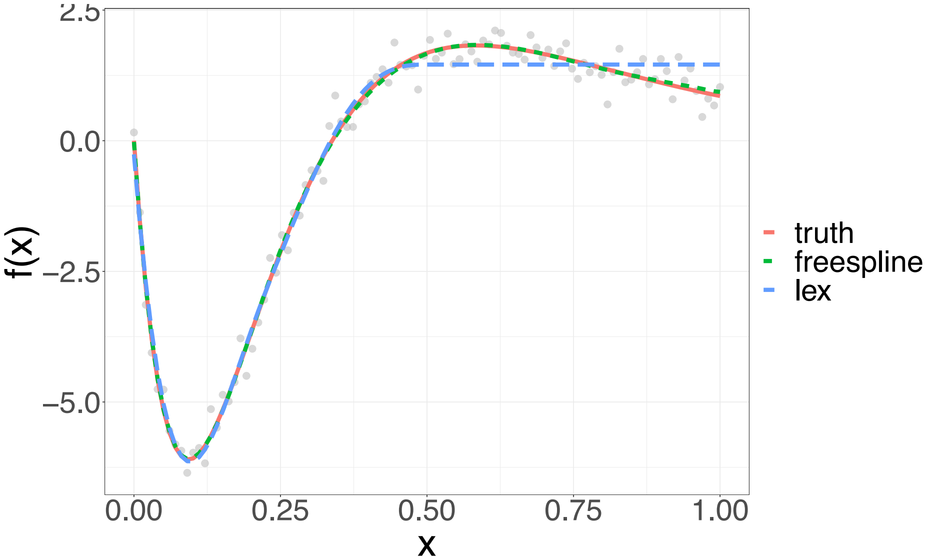


Figure 58: Nonparametric regression fit to noisy measurements of the function $f(x) = 26\exp(-3.25x) - 4\exp(-6.5x) + 3\exp(-9.75x)$ using the freespline algorithm with a Dirichlet process prior on the probability density determining the knot locations. Figure reproduced from Edwards & Gair (2020).

represented by a random cumulative density function, H, defined on the interval $[0, 1]$, with the j'th of $k - r$ internal knots located at $x_j = H(j/(k - r))$. The random density H was assigned a Dirichlet process prior. In Figure 58 we show the result of using this algorithm to fit noisy measurements of a function

$$f(x) = 26\exp(-3.25x) - 4\exp(-6.5x) + 3\exp(-9.75x).$$

We see that the freespline algorithm is able to capture all of the turning points of this function, while another widely used regression algorithm, LEX, is not. In Figure 59 we show another application of that algorithm to obtain a nonparametric fit to the power spectrum of temperature fluctuations in the CMB measured by Planck. The nonparametric fit can be compared to the best fit cosmological model prediction. There is some evidence that the data does not support the up-tick at low multipoles predicted by the model. In fact, there has been extensive debate in the literature about whether the $l = 2, 3$ multipoles are in fact lower than predicted, and these results seem to support that. There is also weak evidence that the data suggests the second and third peaks are further apart than the standard ΛCDM model predicts. Observations of this nature (if they were to be robust in future data sets) would help guide modifications to the model, and this would be much harder without the nonparametric regression tool.

Example: LIGO sky localisation In Del Pozzo et al. (2018), a Dirichlet process Gaussian mixture model (DPGMM) was used to produce a smooth interpolation of the output of LALInference sampling. The aim was to produce a continuous representation of the source localisation volume (sky location and distance), to target electromagnetic follow-up. The Dirichlet process was used as a prior to generate the centres (in 3-dimensions) of Gaussians. The sum of these Gaussians, with weights, was used as a representation of the smooth posterior probability and then constrained by the set of posterior samples

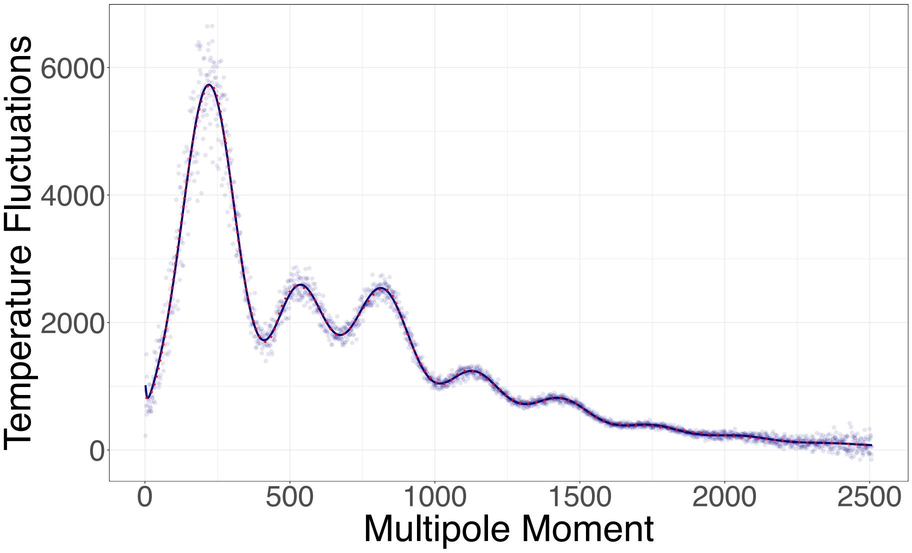


Figure 59: Nonparametric regression fit to the CMB temperature power spectrum, as measured by Planck. The dashed red line is the freespline fit to the data, while the blue line is the prediction of the best fit cosmological model. Figure reproduced from Edwards & Gair (2020).

previously generated by LALInference. In Figure 60 we show the result of this analysis, the distribution of posterior credible volumes computed for a set of injections and using the DPGMM to obtain the credible volumes. This is the only application of Dirichlet processes in a gravitational wave context to date, but they are likely to be powerful tools for fitting nonparametric population models as the number of observations becomes large enough to make this possible.

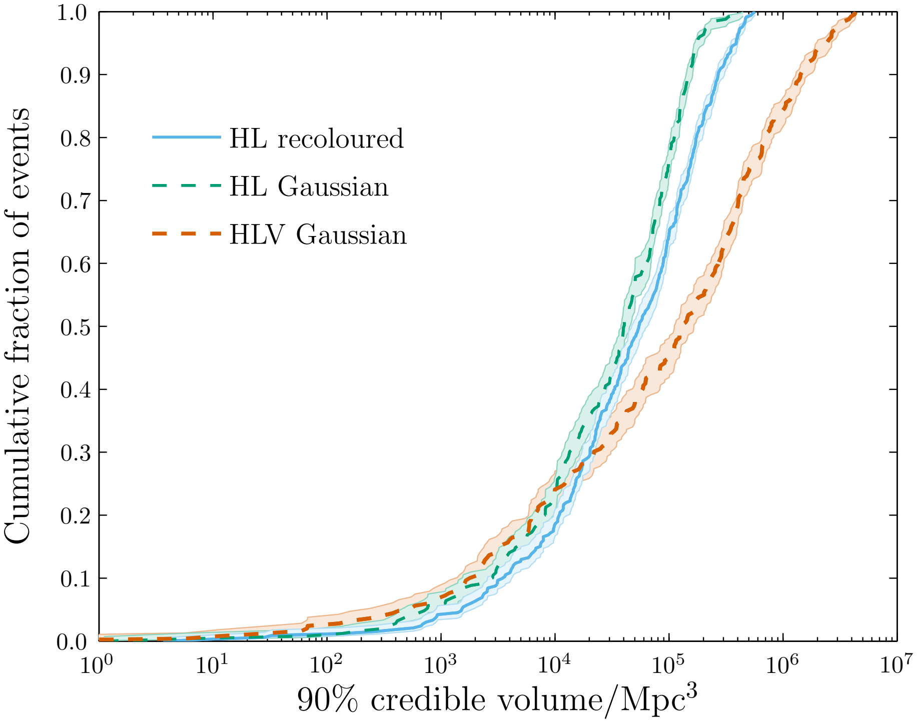


Figure 60: Cumulative distribution function of 90% credible volumes for events observed by the ground-based detector network. The credible volumes were computed by fitting a Dirichlet Process Gaussian Mixture Model to posterior samples generated by LALInference. Figure reproduced from Del Pozzo et al. (2018).