Part IV: Advanced topics in statistics (OPTIONAL)

9 Time Series

We encountered the notion of a time series, or stochastic process, in Section **??** 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

$$\begin{aligned} \mathbb{E}[X_t] &= \mathbb{E}[w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}] \\ &= \mathbb{E}[w_t] + \theta_1 \mathbb{E}[w_{t-1}] + \cdots + \theta_q \mathbb{E}[w_{t-q}] = 0. \end{aligned} \tag{120}$$

Setting $\theta_0 = 1$ the autocovariance is

$$\begin{aligned} \gamma(k) &= \operatorname{cov}(X_t, X_{t+k}) = \mathbb{E}[X_t X_{t+k}] - 0^2 \\ &= \mathbb{E}[(\theta_0 w_t + \cdots + \theta_q w_{t-q})(\theta_0 w_{t+k} + \cdots + \theta_q w_{t+k-q})] \\ &= \sum_{r=0}^{q} \sum_{s=0}^{q} \theta_r \theta_s \mathbb{E}[w_{t-r} w_{t+k-s}]. \end{aligned} \tag{121}$$

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

$$\text{var}(X_t) = \gamma_0 = \sigma^2 \sum_{r=0}^{q} \theta_r^2$$

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

$$\begin{aligned} X_t &= w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q} \\ &= (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q) w_t \\ &= \theta(B) w_t \end{aligned} \tag{122}$$

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

$$\begin{aligned}\gamma(k) = \mathrm{cov}(X_t, X_{t+k}) &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \alpha_1^i w_{t-i}\right)\left(\sum_{j=0}^{\infty} \alpha_1^j w_{t+k-j}\right)\right] \\ &= \sigma^2 \sum_{i=0}^{\infty} \alpha_1^i \alpha_1^{k+i} \text{ for } k \geq 0 \text{ since } \mathbb{E}[w_{t-i} w_{t+k-j}] = 0 \text{ unless } j = k+i \\ &= \frac{\sigma^2 \alpha_1^k}{(1-\alpha_1^2)} \text{ if } |\alpha_1| < 1. \end{aligned} \tag{123}$$

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

$$\begin{aligned} X_t - \alpha_1 X_{t-1} - \alpha_2 X_{t-2} - \ldots - \alpha_p X_{t-p} &= w_t \\ (1 - \alpha_1 B - \alpha_2 B^2 - \ldots - \alpha_p B^p) X_t &= w_t \\ \phi(B) X_t &= w_t. \end{aligned} \tag{124}$$

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

$$\begin{aligned}\gamma(k) &= \mathrm{cov}(X_t, X_{t+k}) \\ &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \psi_i w_{t-i}\right)\left(\sum_{j=0}^{\infty} \psi_j w_{t+k-j}\right)\right] \\ &= \sigma^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k} \text{ for } k \geq 0. \end{aligned} \tag{125}$$

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

$$\begin{aligned} X_t &= \frac{1}{\phi(B)} w_t \\ &= (1 - \alpha_1 B)^{-1} w_t \\ &= (1 + \alpha_1 B + (\alpha_1 B)^2 + \ldots) w_t \\ &= \psi_0 w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} + \ldots \end{aligned} \tag{126}$$

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

$$b \dot{\sim} \text{Bin}(m, 0.05)$$

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

$$\begin{aligned} \gamma(k) &= \text{cov}(X_t, X_{t+k}) = \mathbb{E}[X_t X_{t+k}] - 0 \\ &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \psi_i w_{t-i} \right) \left(\sum_{j=0}^{\infty} \psi_j w_{t+k-j} \right) \right] \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \psi_i \psi_j \mathbb{E}(w_{t-i} w_{t+k-j}). \end{aligned} \tag{127}$$

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

$$\begin{aligned} X_t &= \frac{1 + \beta B}{1 - \alpha B} w_t \\ &= (1 + \alpha B + \alpha^2 B^2 + \ldots)(1 + \beta B) w_t \\ &= [(1 + \alpha B + \alpha^2 B^2 + \ldots) + (\beta B + \beta\alpha B^2 + \beta\alpha^2 B^3 + \ldots)] w_t \\ &= [1 + (\alpha + \beta)B + (\alpha^2 + \alpha\beta)B^2 + (\alpha^3 + \alpha^2\beta)B^3 + \ldots] w_t \\ &= \sum_{i=0}^{\infty} \psi_i w_{t-i} \end{aligned} \tag{128}$$

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

$$\begin{aligned} \text{var}[X_t] &= \sum_{i=0}^{\infty} \psi_i^2 \text{var}[w_{t-i}] = \sigma^2 \sum_{i=0}^{\infty} \psi_i^2 \\ &= [1 + (\alpha + \beta)^2 + (\alpha + \beta)^2\alpha^2 + (\alpha + \beta)^2\alpha^4 + \ldots]\sigma^2 \\ &= \left[1 + \frac{(\alpha + \beta)^2}{(1 - \alpha^2)}\right] \sigma^2. \end{aligned} \tag{129}$$

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

$$\begin{aligned} \psi_0\psi_1 + \psi_1\psi_2 + \psi_2\psi_3 + \ldots &= (\alpha + \beta) + [(\alpha + \beta)^2\alpha + (\alpha + \beta)^2\alpha^3 + \ldots] \\ &= (\alpha + \beta) + \left[\frac{(\alpha + \beta)^2\alpha}{1 - \alpha^2}\right]. \end{aligned} \tag{130}$$

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

$$\begin{aligned} \phi(B) Y_t &= \phi(B)(X_t - \mu) = \phi(B) X_t - \phi(B)\mu \\ &= c + \theta(B) w_T - c = \theta(B) w_t \end{aligned} \tag{131}$$

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

$$\Phi(B)X_t = \theta(B)w_t, \quad \text{where } \Phi(B) = (I - B), \quad \theta(B) = I - \theta B.$$

We can invert this process to obtain

$$\begin{aligned} w_t = \Pi(B)X_t &= \frac{I - B}{I - \theta B}X_t \\ &= (1 - B)(1 + \theta B + \theta^2 B^2 + \ldots)X_t \\ &= [1 - (1 - \theta)B - (1 - \theta)\theta B^2 - (1 - \theta)\theta^2 B^3 + \ldots]X_t \\ &= \sum_{i=0}^{\infty} \pi_i X_{t-i} \end{aligned} \tag{132}$$

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

$$||g||_2^2 \stackrel{def}{=} \int_0^1 [g(x)]^2 dx.$$

   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

$$\text{supp}(g) = \{x : g(x) \neq 0\}.$$

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 $K(x) = \frac{\pi}{4}\cos(\pi x/2)\mathbf{1}(|x| \le 1)$.

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* $m \ge 1$.

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* $K(x) = \frac{1}{2}\mathbf{1}(|x| \le 1)$ *is a kernel then* $K(x) = \frac{1}{4}\mathbf{1}(|x| \le 2)$ *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:

$$\widehat{p}_n(x) = \frac{1}{n}\sum_{i=1}^n K_h(x_i - x), \quad \widehat{p}_n(x,y) = \frac{1}{n}\sum_{i=1}^n K_h(x_i - x)K_h(y_i - y). \tag{133}$$

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 (133).

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:

$$\int_{-\infty}^{+\infty} xK(x)dx = 0, \quad 0 < \mu_2(K) \stackrel{def}{=} \int_{-\infty}^{+\infty} x^2 K(x)dx < \infty,$$

$$||K||_2^2 = \int_{-\infty}^{+\infty} [K(x)]^2 dx < \infty.$$

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

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x) \approx \int_0^1 K_h(z-x)dz &= \int_0^1 K\left(\frac{z-x}{h}\right) d\left(\frac{z}{h}\right) = \int_{\frac{0-x}{h}\to}^{\frac{1-x}{h}} K(v)dv \\ &\approx \int_{-\infty}^{+\infty} K(v)dv = 1\end{aligned}$$

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: $v(x) \approx \frac{\sigma^2}{nh}||K||_2^2$:

$$\begin{aligned}
v(x) = \sigma^2 \sum_{i=1}^{n} [w_i(x)]^2 &= \sigma^2 \sum_{i=1}^{n} \frac{[K_h(X_i - x)]^2}{\left[\sum_{j=1}^{n} K_h(X_j - x)\right]^2} \\
&\approx_{\left\{\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x) \approx 1\right\}} \frac{\sigma^2}{n^2} \sum_{i=1}^{n} [K_h(X_i - x)]^2 \\
\left\{\frac{1}{n}\sum_{i=1}^{n} \to \int_0^1\right\} &\approx \frac{\sigma^2}{n} \int_0^1 [K_h(z - x)]^2 dz = \frac{\sigma^2}{nh} \int_0^1 \left[K\left(\frac{z-x}{h}\right)\right]^2 d\left(\frac{z-x}{h}\right) \\
\left\{v = \frac{z-x}{h}\right\} &= \frac{\sigma^2}{nh} \int_{-x/h}^{(1-x)/h} [K(v)]^2 dv \approx \frac{\sigma^2}{nh} \int_{-\infty}^{\infty} [K(v)]^2 dv \\
&= \frac{\sigma^2}{nh} ||K||_2^2.
\end{aligned}$$

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, $||K_h||_2^2 = \mu_2(K_h)^2 \Leftrightarrow \frac{1}{h}||K||_2^2 = \left[h^2\mu_2(K)\right]^2$ 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:

$$\int K(x)dx = 1, \quad \int xK(x)dx = 0, \quad \int x^2K(x)dx = 1.$$

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:

$$\text{bias}(\widehat{f^{NW}}(x)) = \mathbb{E}(\widehat{f^{NW}}(x)) - f(x) = \sum_{i=1}^{n} W_i^{NW}(x)\left[f(x_i) - f(x)\right].$$

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,

$$\begin{aligned}
|\text{bias}(\widehat{f}^{NW}(x))| &= \frac{\left|\sum_i K(\frac{x_i - x}{h})(f(x_i) - f(x))\right|}{\sum_i K(\frac{x_i - x}{h})} = \frac{\left|\sum_{i:\,|x_i - x| \leq h} K(\frac{x_i - x}{h})[f(x_i) - f(x)]\right|}{\sum_i K(\frac{x_i - x}{h})} \\
&\leq \frac{\sum_{i\,:|x_i - x| \leq h} K(\frac{x_i - x}{h})\,|f(x_i) - f(x)|}{\sum_i K(\frac{x_i - x}{h})} \leq \frac{\sum_{i\,:|x_i - x| \leq h} K(\frac{x_i - x}{h})\, M|x_i - x|^\beta}{\sum_i K(\frac{x_i - x}{h})} \\
&\leq Mh^\beta,
\end{aligned}$$

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,

$$\begin{aligned}
v(x) &= \sigma^2 \sum_{i=1}^n \frac{[K_h(X_i - x)]^2}{\left[\sum_{j=1}^n K_h(X_j - x)\right]^2} && \text{assumption d) } K(z) \geq 0 \text{ for all } z \\
&\leqslant \sigma^2 \frac{\frac{K_{\max}}{h}\sum_{i=1}^n K_h(X_i - x)}{\left[\sum_{j=1}^n K_h(X_j - x)\right]^2} && \text{assumption d), } \forall u, K(u) \leqslant K_{\max} \text{ implies } K_h(X_i - x) = \tfrac{1}{h}K\left(\tfrac{X_i - x}{h}\right) \leqslant \tfrac{K_{\max}}{h} \\
&\leqslant \sigma^2 \frac{\frac{K_{\max}}{h}\sum_{i=1}^n K_h(X_i - x)}{n\lambda_0 \sum_{j=1}^n K_h(X_j - x)} && \text{assumption c) } \exists \lambda_0 > 0 \text{ such that } \forall x \in [0,1], \\
&= \frac{\sigma^2 K_{\max}}{nh\lambda_0} && \sum_{i=1}^n K_h(X_i - x) \geqslant n\lambda_0.
\end{aligned}$$

□

Now we consider the bounds on the MSE of the NW estimator. Under the conditions of Proposition 10.2,

$$\mathrm{MSE}(\widehat{f}_n^{NW}(x_0)) = [\mathrm{bias}(\widehat{f}_n^{NW}(x_0))]^2 + \mathrm{Var}(\widehat{f}_n^{NW}(x_0)) \le M^2 h^{2\beta} + \frac{\sigma^2 K_{\max}}{nh\lambda_0}.$$

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

$$\begin{aligned}\mathrm{MSE}(\widehat{f}_{n,h_n}^{NW}(x_0)) &\le M^2\left(\frac{\sigma^2 K_{\max}}{2\beta M^2\lambda_0 n}\right)^{2\beta/(2\beta+1)} + \frac{\sigma^2 K_{\max}}{n\lambda_0}\left(\frac{2\beta M^2\lambda_0 n}{\sigma^2 K_{\max}}\right)^{1/(2\beta+1)} \\ &\le (1+2\beta)M^{2/(2\beta+1)}\left(\frac{\sigma^2 K_{\max}}{2\beta\lambda_0 n}\right)^{2\beta/(2\beta+1)} \to 0 \quad \text{as} \quad n\to\infty.\end{aligned}$$

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 $\frac{1}{2n}\sum_{i:|x_i-x|\le h}\frac{1}{h} \ge \lambda_0$, $h \ge 1/2n$.

Let's count the number of integers i between 1 and n such that $|i/n - x| \le h$. Since

$$|i/n - x| \le h \Leftrightarrow (nx - nh) \le i \le (nx + nh),$$

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 $\lfloor 2nh\rfloor \le 2nh < \lfloor 2nh\rfloor + 1$ 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

$$\frac{1}{2n}\sum_{i:|x_i-x|\le h}\frac{1}{h} \ge \frac{2nh-1}{2nh} = 1 - 1/(2nh) = \lambda_0$$

Assumption d) is satisfied with $K_{\max} = 1/2$.

Therefore, by Proposition 10.2, for $n = 12$ and $h > 1/24$,

$$|b(x)| \le Mh^\beta = 5\sqrt{h}, \quad v(x) \le \frac{1}{2nh(1-1/(2nh))} = \frac{1}{2nh-1} = \frac{1}{24h-1}.$$

The corresponding MSE (and MISE) for $\hat{f}^{NW}(x)$ is bounded by

$$MSE(\hat{f}^{NW}(x)) = b^2(x) + v(x) \le 25h + \frac{1}{24h-1}.$$

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, $\forall f \in H^\beta(M), \forall x \in [0,1]$,

$$\mathbb{E}\left[\left(\widehat{f}_n^{\,NW}(x) - f(x)\right)^2\right] \leq Cn^{\frac{-2\beta}{2\beta+1}}$$

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

$$\begin{aligned} \theta_x &= \left(f(x), f'(x)h, f''(x)h^2, \ \ldots \ , f^{(k)}(x)h^k \right)^T \\ U_{x,i} &= \left(1, \frac{x_i - x}{h}, \frac{1}{2!} \left(\frac{x_i - x}{h} \right)^2, \ldots, \frac{1}{k!} \left(\frac{x_i - x}{h} \right)^k \right)^T \end{aligned}$$

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 $\widehat{\theta}(x) = \left(\widehat{\theta}_0(x), \widehat{\theta}_1(x), \ldots, \widehat{\theta}_k(x) \right)^T$ 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 $m = 1, \ldots, k$, $\widehat{\theta}_m(x)/h^m$ 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

$$\begin{aligned} \widehat{\theta}_x &= \arg \min_{\theta_x} \left\{ \sum_{i=1}^{n} (Y_i - U_{x,i}^T \theta_x)^2 K \left(\frac{X_i - x}{h} \right) \right\} \\ &= \arg \min_{\theta_x} \left\{ \sum_{i=1}^{n} (\theta_x^T U_{x,i} U_{x,i}^T \theta_x - 2 U_{x,i}^T \theta_x Y_i + Y_i^2) K \left(\frac{X_i - x}{h} \right) \right\} \\ &= \arg \min_{\theta_x} \left\{ \theta_x^T \cdot \sum_{i=1}^{n} U_{x,i} U_{x,i}^T K \left(\frac{X_i - x}{h} \right) \cdot \theta_x - \theta_x^T \cdot 2 \sum_{i=1}^{n} Y_i U_{x,i} K \left(\frac{X_i - x}{h} \right) + C_{X_i, Y_i}(x) \right\} \end{aligned}$$

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**:

$$f_n^{LP}(x) = e_1^T B^{-1}(x)a(x) = e_1^T B^{-1}(x) \cdot \frac{1}{n}\sum_{i=1}^{n} Y_i U_{x,i} K_h\left(X_i - x\right)$$
$$= \sum_{i=1}^{n} Y_i \cdot \frac{1}{n} K_h\left(X_i - x\right) \sum_{j=0}^{k} [B^{-1}(x)]_{0,j} \frac{1}{j!}\left(\frac{x_i - x}{h}\right)^j$$
$$= \sum_{i=1}^{n} Y_i w_i(x)$$

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

$$\lambda_{\min}(B(x)) \geqslant \lambda_0 \quad , \text{ where } B(x) = \frac{1}{n}\sum_{i=1}^{n} U_{x,i}U_{x,i}^T K_h(X_i - x);$$

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

$$\text{MISE}\left(\widehat{f}^{LP}(x)\right) = \int_0^1 \text{MSE}(\widehat{f}^{LP}(x))dx \leqslant C n^{-\frac{2\beta}{2\beta+1}}$$

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*

$$\begin{aligned} \limsup_{n\to\infty} \sup_{f\in\mathbb{H}^{\beta}(M)} \sup_{x_0\in[0,1]} \mathbb{E}\left[n^{\frac{\beta}{2\beta+1}} |f(x_0) - \widehat{f}_n(x_0)|\right]^2 &\leq C < \infty, \\ \limsup_{n\to\infty} \sup_{f\in\mathbb{H}^{\beta}(M)} \mathbb{E}\left[n^{\frac{\beta}{2\beta+1}} ||f - \widehat{f}_n||_2\right]^2 &\leq C < \infty, \end{aligned}$$

where C is a constant depending only on β, M, a_0, λ_0, $\sigma^2_{\max}$, $K_{\max}$ and α.

10.3 Smoothing Splines

10.3.1 Definition

Definition 10.16. *A smoothing spline is the penalised least squares estimator of f:*

$$\widehat{f}_n^{\text{pen}}(x) = \arg\min_{f\in\mathcal{C}^2}\left[\sum_{i=1}^{n}(Y_i - f(x_i))^2 + \lambda\,\text{pen}(f)\right] \tag{134}$$

with penalty function $\text{pen}(f) = \int[f''(x)]^2dx = ||f''||_2^2$*; $\lambda > 0$ 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 $a \le t_1 < .. < t_N \le b$ 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*

$$\begin{aligned} h_1(x) &= 1, h_2(x) = x, \\ h_j(x) &= (x - t_{j-2})_+^3 - \frac{(t_N - t_{j-2})}{(t_N - t_{N-1})}(x - t_{N-1})_+^3 \\ &\qquad + \frac{(t_{N-1} - t_{j-2})}{(t_N - t_{N-1})}(x - t_N)_+^3, \quad \forall\ 3 \leqslant j \leqslant N, \end{aligned}$$

where $(x-y)_+^3 = \max\left\{(x-y)^3, 0\right\}$

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 $h_j(x)$, $j = 1, 2, \ldots, N$. Hence, minimising over functions f

$$\begin{aligned}\widehat{f}_{n,\lambda}^{SS} &= \arg\min_{f\in C^2}\left\{\sum_{i=1}^{N}\left(Y_i - f(x_i)\right)^2 + \lambda\int\left[f''(x)\right]^2 dx\right\}\\ &= \arg\min_{f\in C^2}\left\{\sum_{i=1}^{N}\left(f(x_i)^2 - 2f(x_i)Y_i + Y_i^2\right) + \lambda\int\left[f''(x)\right]^2 dx\right\}\end{aligned}$$

is equivalent to minimising the following expression over the $(n+2)$-dimensional vector β:

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^N}\left\{\sum_{i=1}^{N}\left[\sum_{j=1}^{N}\beta_j h_j(x_i)\right]^2 - 2\sum_{i=1}^{N}\left[\sum_{j=1}^{N}\beta_j h_j(x_i)\right]Y_i + \lambda\int\left[\sum_{j=1}^{N}\beta_j h_j''(x)\right]^2 dx\right\}$$

$$= \arg\min_{\beta\in\mathbb{R}^N}\left\{\beta^T H^T H\beta - 2\beta^T H^T Y + \lambda\beta^T\Omega\beta\right\},$$

where $N\times N$ matrix H has entries $H_{ij} = h_j(x_i)$, $i = 1,\ldots,N$, $j = 1,\ldots,N$, and $N\times N$ matrix Ω has elements $\Omega_{j\ell} = \int h_j''(x)h_\ell''(x)dx$, $j,\ell = 1,\ldots,N$.

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 $\widehat{\beta} = (\widehat{\beta}_1,\ldots,\widehat{\beta}_N)^T$ *is given by*

$$\widehat{\beta} = (H^T H + \lambda\Omega)^{-1}H^T Y$$

where $Y = (Y_1,\ldots,Y_n)^T$, *and matrices* $H = (H_{ij})$ *and* $\Omega = (\Omega_{jl})$ *have entries*

$$H_{ij} = h_j(x_i),\quad \Omega_{jl} = \int_a^b h_j''(x)h_l''(x)dx,\quad i\in 1,\ldots,n,\quad j,l\in 1,\ldots,N$$

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

$$w_i(x) = \sum_{j=1}^{N} h_j(x)\left[\left(H^T H + \lambda\Omega\right)^{-1} H^T\right]_{ji}$$

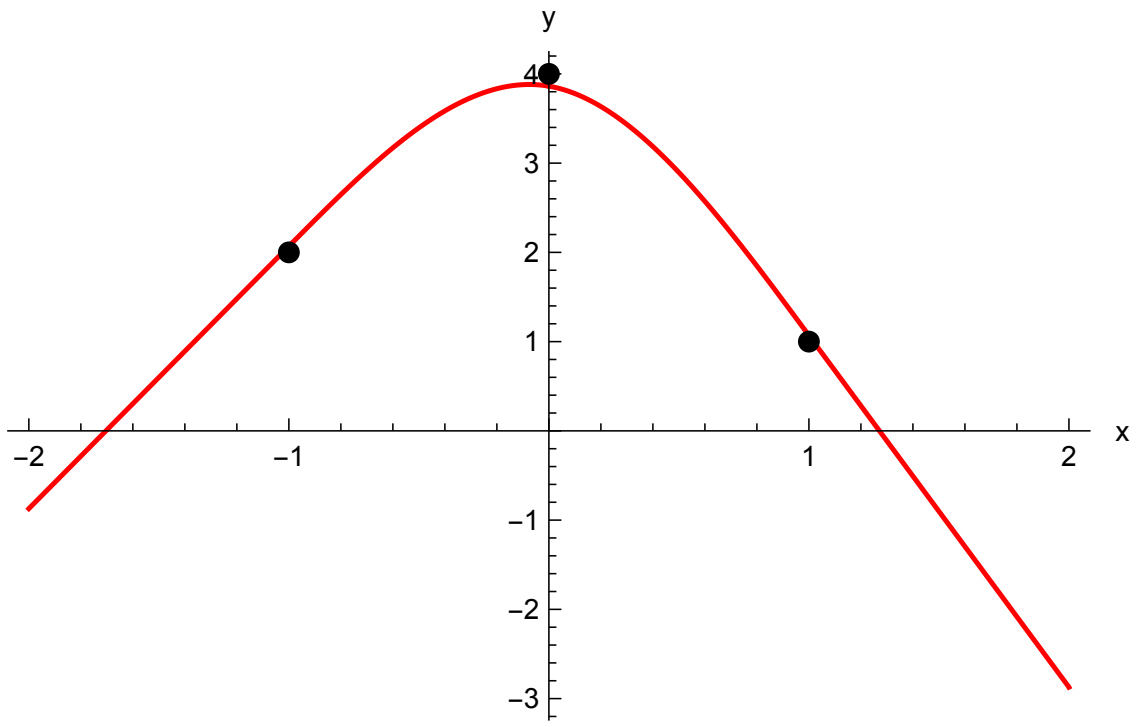


Figure 46: Smoothing spline for example 10.7.

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*

$$\widehat{f}_n^{SS}(x) = \sum_{i=1}^{N}\sum_{j=1}^{N}[(H^TH + \lambda\Omega)^{-1}H^T]_{ji}h_j(x)Y_i.$$

The matrices necessary for the calculation are $H = (H_{ij})$, $H_{ij} = h_j(x_i)$:

$$H = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 6 \end{pmatrix}, \quad H^TH = \begin{pmatrix} 3 & 0 & 7 \\ 0 & 2 & 6 \\ 7 & 6 & 37 \end{pmatrix}$$

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 46.*

10.3.2 Choice of Regularisation Parameter λ

In papplications, λ is usually chosen using cross-validation

$$\widehat{\lambda} = \arg\min_{\lambda>0}\left\{\sum_{i=1}^{n}\left(Y_i - \widehat{f}_{\lambda,-i}(x_i)\right)^2\right\}$$

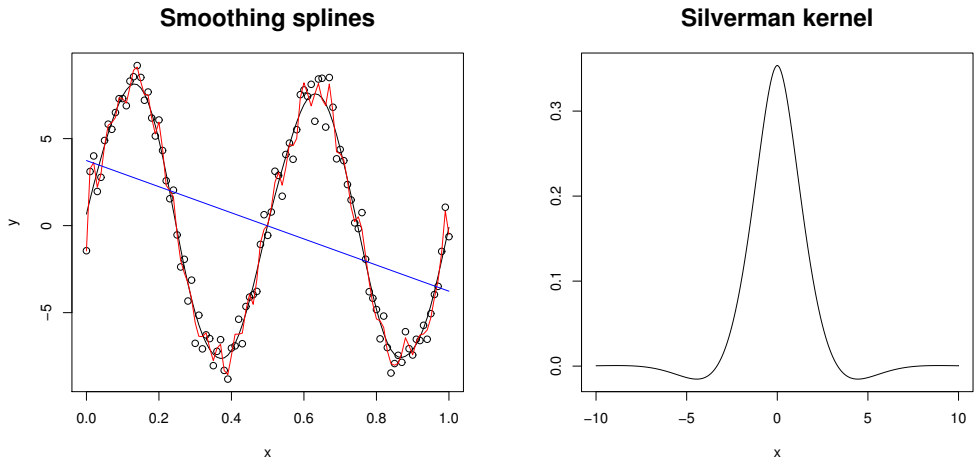


Figure 47: Left: smoothing spline estimator Right: Silverman kernel

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 47 (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 47 (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

$$\begin{aligned} \theta_{\mathbf{x}} &= (\theta^0, \theta_1^1, \theta_2^1, \ldots, \theta_m^1, \theta_{11}^2, \theta_{12}^2, \ldots \theta_{mm}^2, \ldots, \theta_{mm\cdots m}^k) \\ U_{\mathbf{x},i} &= \left(1, \frac{x_{1i} - x_1}{h}, \frac{x_{2i} - x_2}{h}, \ldots, \frac{x_{mi} - x_m}{h}, \frac{1}{2!}\left(\frac{x_{1i} - x_1}{h}\right)^2, \left(\frac{x_{1i} - x_1}{h}\right)\left(\frac{x_{2i} - x_2}{h}\right), \right. \\ &\qquad \left. \ldots, \frac{1}{2!}\left(\frac{x_{mi} - x_m}{h}\right)^2, \ldots, \frac{1}{k!}\left(\frac{x_{mi} - x_m}{h}\right)^k\right). \end{aligned}$$

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*

$$\hat{\beta}_2^T = (1.51025, 0.941748, -0.647249).$$

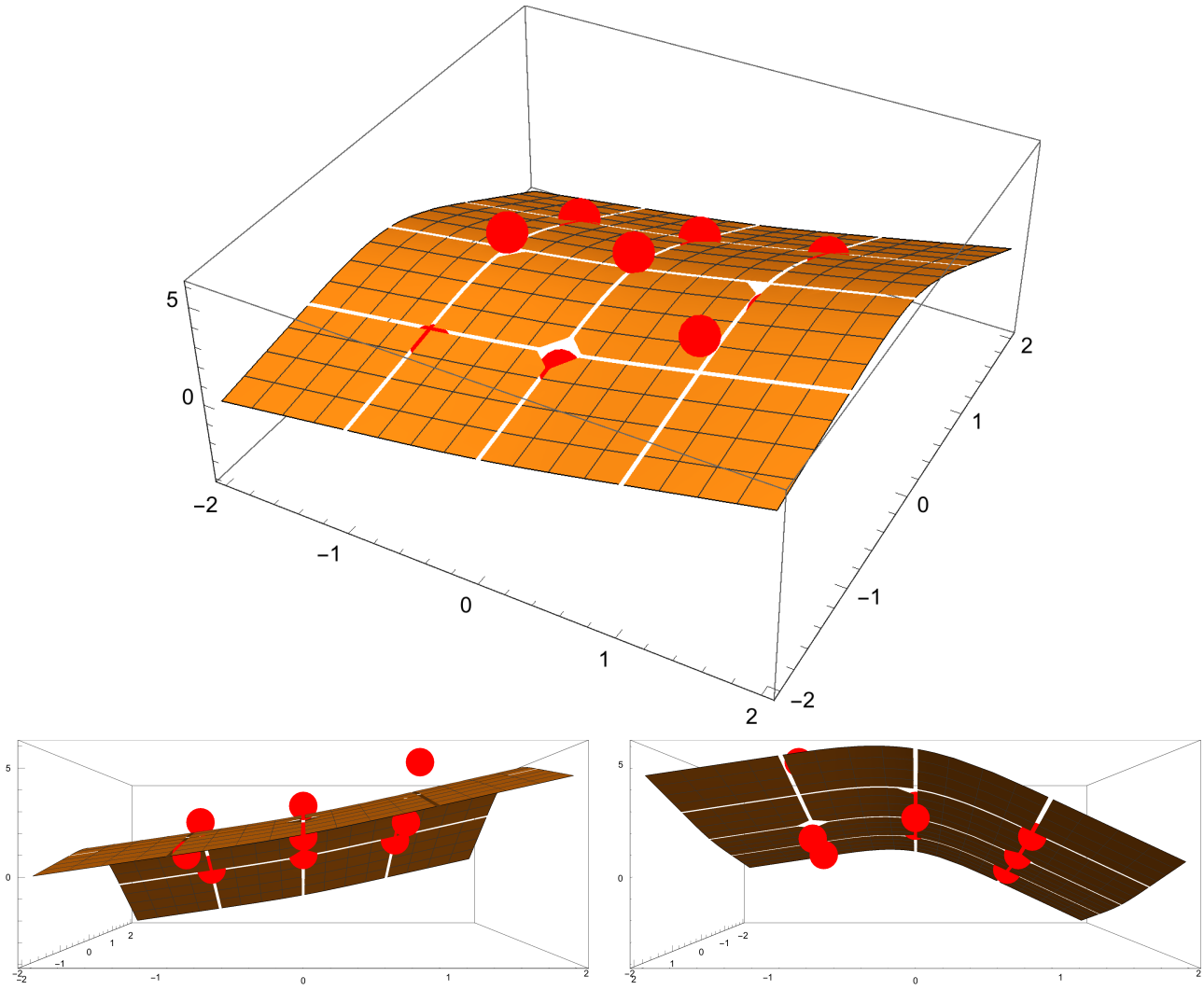


Figure 48: Data (red points) and general additive model fit (shaded surface) for example 10.8. The top plot shows the full surface, while the bottom two plots show the surface from the x_1 and x_2 sides respectively.

The algorithm has now converged and we obtain our general additive model estimate of $f(x_1, x_2)$ *as*

$$\hat{f}(x_1, x_2) = \frac{22}{9} + \sum_{i=1}^{3} \beta_{1i} h_i(x_1) + \sum_{i=1}^{3} \beta_{2i} h_i(x_2)$$

$$\textit{where } h_1(x) = 1, \qquad h_2(x) = x, \qquad h_3(x) = (x+1)_+^3 - 2(x)_+^3 + (x-1)_+^3.$$

The raw data and the GAM estimate are shown in Figure 48.

10.5 Wavelet Estimators

We return again to the nonparametric regression model

$$Y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \qquad \mathbb{E}(\varepsilon_i) = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma^2,\ \text{independently.}$$

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: $\varphi_{2k}(x) = 1$, $\varphi_{2k}(x) = \cos(2\pi kx)$, $\varphi_{2k+1}(x) = \sin(2\pi kx)$, $k = 1, 2, \ldots$, $x \in [0,1]$ (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.

(1) for large N, approximate $f(x) \approx \sum_{k=0}^{N} \theta_k \varphi_k(x)$

(2) construct an estimator $\widehat{\theta}_k$ of θ_k from data $(y_1, \ldots, y_n)$, $k = 0, 1, \ldots, N$

(3) plug in the estimator $\widehat{\theta}_k$ 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

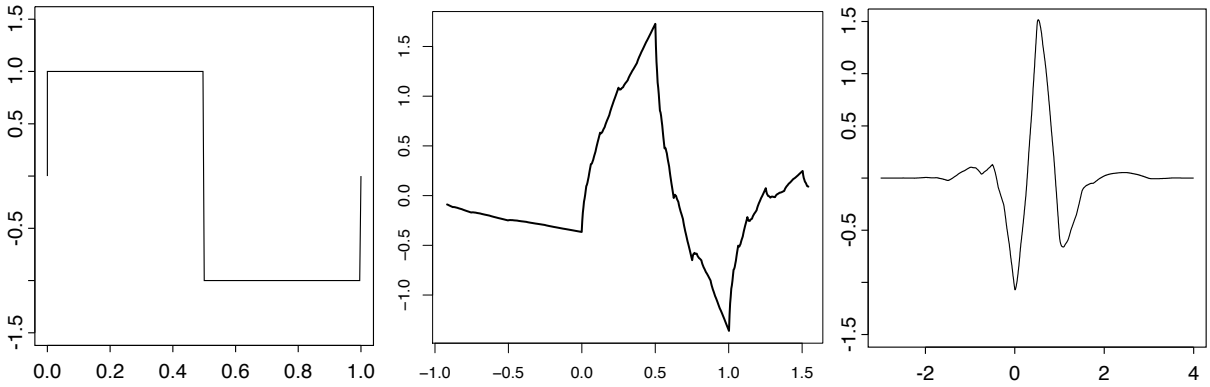
A wavelet basis is constructed using two functions, a scaling function $\phi(x)$ and a wavelet function $\psi(x)$ that are also called the father and mother wavelet respectively. They satisfy the following properties:

$$\int \phi(x) dx = 1, \quad \int \psi(x) dx = 0.$$

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 49: 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 49, and the structure of the wavelet basis is illustrated in Figure 50.

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*

$$\int \phi(x)dx = 1, \quad \int \psi(x)dx = 0, \quad \int \psi_{jk}(x)dx = 0.$$

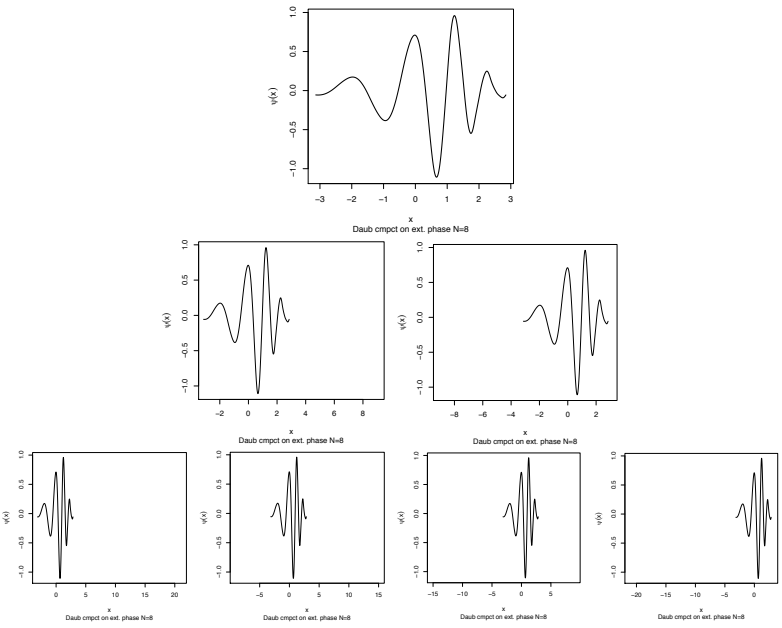
Check that the basis $\{\phi, \psi_{jk}, j = 0, 1, \ldots, k = 0, \ldots, 2^j - 1\}$ defined by these functions is orthonormal, that is, that the functions are normalised

$$||\phi||_2^2 = \int \phi^2(x)dx = 1, \qquad ||\psi||_2^2 = \int \psi^2(x)dx = 1, \qquad ||\psi_{jk}||_2^2 = \int \psi_{jk}^2(x)dx = 1,$$

and are orthogonal:

$$\int \phi(x)\psi_{jk}(x)dx = 0, \qquad \int \psi_{jk}(x)\psi_{\ell m}(x) = 0 \text{ for } (j,k) \neq (\ell, m).$$

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 50: 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. $V_0 \subset V_1 \subset V_2 \subset \ldots$, $\quad \mathrm{Clos}(\bigcup_j V_j) = L^2(\mathbb{R})$, *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*:*

$$V_0 = \left\{ g \in L^2(\mathbb{R}) : \quad g(x) = \sum_{k\in\mathbb{Z}} c_k \phi(x-k) \text{ for some } (c_k)_{k\in\mathbb{Z}} \right\},$$

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

$$\phi(x) = \sqrt{2} \sum_{k\in\mathbb{Z}} h_k \phi(2x - k). \tag{135}$$

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}: $|\tilde{\theta}_{jk} - \theta_{jk}| \le C/n$. 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 $|\tilde{\theta}_{jk} - \theta_{jk}| \le C/n$ 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, $|\tilde{\theta}_{jk} - \theta_{jk}| \le C/n$ 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

$$\begin{aligned} d_{jk} &= w_{jk} + \varepsilon_{jk}, \quad 0 \le j \le J-1,\ k = 0, \ldots, 2^j - 1, \\ c_{00} &= u_{00} + \varepsilon_0, \end{aligned}$$

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 $f(x_i)$, $x_i = i/n$, $i = 1, \ldots, n$. 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 $u_{J,k} = f(x_{k+1})$, $k = 0, 1, \ldots, 2^J - 1 = n - 1$. 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 $0 \le j \le J - 1$, $\quad k = 0, \ldots, 2^j - 1$.

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 $c_{J0} = Y_1, \ldots, c_{J,2^J-1} = Y_n$, 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:

$$\widehat{w}_{jk} = d_{jk},\ j \le J_0 - 1; \quad \widehat{w}_{jk} = 0,\ j \ge J_0; \quad \widehat{u}_{00} = c_{00}.$$

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 $j \le J_0 - 1$ and $\widehat{w}_{jk} = 0$ for $j > J_0$.

10.5.10 Thresholding Estimators for threshold λ

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

$$\widehat{w}_{jk} = \begin{cases} d_{jk} - \lambda, & d_{jk} > \lambda \\ 0, & -\lambda \le d_{jk} \le \lambda \\ d_{jk} + \lambda, & d_{jk} < -\lambda \end{cases}$$

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^2_m distribution under the null hypothesis where m is the number of coefficients tested to be zero, that is, $m = \text{Card}\{(j,k) : a < \frac{k+1/2}{2^j} < b,\ 0 \le j < J,\ 0 \le k \le 2^j - 1\}$.

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* $T = d_{21}^2/\sigma^2 \sim \chi^2_1$*, therefore we reject* H_0 *at a 5% significance level if* $T = d_{21}^2/\sigma^2 > \chi^2_1(5\%) = 3.841$*. 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 $T = (d_{11}^2 + d_{22}^2 + d_{23}^2)/\sigma^2 \sim \chi^2_3$ *under* H_0*. That is, we reject the null hypothesis at a 5% significance level if* $T > \chi^2_3(5\%) = 7.815$*. 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

$$\hat{f}_{J_0}(x) = \sum_{i=1}^{n}Y_iW_i(x),$$

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 $j \leq J_0 - 1$,

$$\begin{aligned}
\mathrm{Var}\left(\hat{\theta}_{jk}\right) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(x_i)Y_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\psi_{jk}^2(x_i)\mathrm{Var}\left(Y_i\right) = \frac{\sigma^2}{n}\,\frac{1}{n}\sum_{i=1}^{n}\psi_{jk}^2(x_i) \\
&= \frac{\sigma^2}{n}(1+o(1)),
\end{aligned}$$

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 $j \leq J_0 - 1$, 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. $|f(x)| \leq C_f$ for all $x \in [0,1]$. We assume that the wavelet function ψ is such that $|\psi(x) - \psi(y)| \leq M_\psi|x-y|$ 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). 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

$$\begin{aligned}
|\mathbb{E}\left(\hat{\theta}_{jk} - \theta_{jk}\right)| &\leq \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} |f(x)\psi_{jk}(x) - f(x_i)\psi_{jk}(x_i)|\,dx \\
&\leq \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} \left[|f(x)\psi_{jk}(x) - f(x)\psi_{jk}(x_i)| + |f(x)\psi_{jk}(x_i) - f(x_i)\psi_{jk}(x_i)|\right]dx \\
&\leq \max_x |f(x)|2^{j/2}\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} \left|\psi(2^j x - k) - \psi(2^j x_i - k)\right|dx \\
&\quad + \sum_{i=1}^{n}|\psi_{jk}(x_i)|\int_{x_{i-1}}^{x_i}|f(x) - f(x_i)|\,dx.
\end{aligned}$$

Considering the first term on the right hand side, we have

$$\begin{aligned}
\int_{x_{i-1}}^{x_i} \left|\psi(2^j x - k) - \psi(2^j x_i - k)\right|dx &\leq M_\psi \int_{x_{i-1}}^{x_i} |2^j x - k - (2^j x_i - k)|dx \\
&\leq 0.5 M_\psi 2^j n^{-2}.
\end{aligned}$$

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. $k2^{J-j}+1 \le i <\le (k+1)2^{J-j}$. There are 2^{J-j} of such i. Thus,

$$\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}|\psi_{jk}(x)-\psi_{jk}(x_i)|dx \le 0.5M_\psi 2^j n^{-2}2^{J-j} = 0.5M_\psi n^{-2}2^J = 0.5M_\psi n^{-1},$$

using $n=2^J$ and hence

$$\max_x|f(x)|2^{j/2}\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left|\psi(2^jx-k)-\psi(2^jx_i-k)\right|dx \le 0.5C_fM_\psi 2^{j/2}n^{-1}.$$

For the second term, we have

$$\int_{x_{i-1}}^{x_i}|f(x)-f(x_i)|\,dx \;\;\le\;\; M_f\int_{x_{i-1}}^{x_i}|x-x_i|^\beta \le \frac{M_f}{(\beta+1)n^{\beta+1}},$$

and using the restriction to the support of ψ_{jk}

$$|\psi_{jk}(x_i)| \le 2^{j/2}C_\psi \mathbf{1}(k2^{J-j}+1 < i < (k+1)2^{J-j}),$$
$$\Rightarrow \sum_{i=1}^{n}|\psi_{jk}(x_i)| \le 2^{j/2}C_\psi\sum_{i=1}^{n}\mathbf{1}(k2^{J-j}+1 \le i \le (k+1)2^{J-j}) \le 2^{J-j/2}C_\psi \le C_\psi n2^{-j/2}.$$

Thus,

$$|\mathbb{E}\hat\theta_{jk}-\theta_{jk}| \;\;\le\;\; 0.5C_fM_\psi 2^{j/2}n^{-1} + \frac{M_fC_\psi}{(\beta+1)}2^{-j/2}n^{-\beta}$$

again using $n=2^J$ and $j<J$.

MSE $(\hat\theta_{jk})$ for $j \ge J_0$

For $j \ge J_0$, $\hat\theta_{jk}=0$, and therefore the MSE $(\hat\theta_{jk}) = \mathbb{E}(\hat\theta_{jk}-\theta_{jk})^2 = \theta_{jk}^2$.
For $f \in H^\beta(M_f)$, $|\theta_{jk}| \le M_f 2^{-j(\beta+1/2)}$ 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 $|f(x)| \le C_f$ for all $x \in [0,1]$;*
- *ψ is such that supp$(\psi) \subseteq [0,1]$, $|\psi(x)-\psi(y)| \le M_\psi|x-y|$ 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).*

Then, for $\hat\theta_{jk} = \frac{1}{n}\sum_{i=1}^n \psi_{jk}(x_i)Y_i$,

$$\begin{aligned} Var\left(\hat\theta_{jk}\right) &= \frac{\sigma^2}{n}(1+o(1)) \quad \text{as } n\to\infty, \\ |\text{bias}(\hat\theta_{\text{jk}})| &\le c_1 2^{j/2}n^{-1} + c_2 2^{-j/2}n^{-\beta}, \end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}||f-\hat{f}_{J_0}||_2^2 &= \mathbb{E}(\theta_0-\hat{\theta}_0)^2+\sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1}\mathbb{E}(\hat{\theta}_{jk}-\theta_{jk})^2+\sum_{j=J_0}^{\infty}\sum_{k=0}^{2^j-1}\theta_{jk}^2 \\
&\leq 2^{J_0}\frac{\sigma^2}{n}(1+o(1))+2c_1^2n^{-2}[1+\sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1}2^j] \\
&\quad +2c_2^2n^{-2\beta}[1+\sum_{j=0}^{J_0-1}\sum_{k=0}^{2^j-1}2^{-j}]+M_f^2\sum_{j=J_0}^{\infty}\sum_{k=0}^{2^j-1}2^{-j(2\beta+1)} \\
&= 2^{J_0}\frac{\sigma^2}{n}(1+o(1))+2c_1^2n^{-2}(2^{2J_0}+2)/3+2c_2^2n^{-2\beta}(J_0+1)+M_f^2\frac{2^{-2\beta J_0}}{1-2^{-2\beta}} \\
&\leq \sigma^2\frac{N}{n}(1+o(1))+\tilde{c}_1n^{-2}N^2+\tilde{c}_2n^{-2\beta}\log n+\tilde{c}_3N^{-2\beta}+\tilde{c}_4n^{-2}
\end{aligned}$$

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:

$$MISE(\hat{f}_{J_0}) \leq \sigma^2\frac{N}{n}(1+o(1))+\tilde{c}_3N^{-2\beta}(1+o(1))$$

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 $\mathcal{D}$,

$$[y(\mathbf{x}_i)] \sim N(\boldsymbol{m},\, \boldsymbol{K})\,, \tag{136}$$

where the mean vector and covariance matrix of this Gaussian distribution are fixed by the corresponding functions of the GP,

$$[\boldsymbol{m}]_i = m(\mathbf{x}_i)\,, \quad [\boldsymbol{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)\,, \tag{137}$$

with probability density function

$$P\left(\{y(\mathbf{x}_i\}\right) = \frac{1}{\sqrt{(2\pi)^N\,|\boldsymbol{K}|}} \exp\left(-\frac{1}{2}\sum_{i,\,j}(y(\mathbf{x}_i) - m(\mathbf{x}_i))\left[\boldsymbol{K}^{-1}\right]_{ij}(y(\mathbf{x}_j) - m(\mathbf{x}_j))\right)\,. \tag{138}$$

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)

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left[-\frac{1}{2} g_{ab}(\mathbf{x}_i - \mathbf{x}_j)^a (\mathbf{x}_i - \mathbf{x}_j)^b\right], \tag{139}$$

which defines a stationary, smooth GP. In Eq. (139), 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.

When the set of points, $\{\mathbf{x}_i\}$ coincides with the training set, the probability in Eq. (138) 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 the specified mean and 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

$$\begin{aligned} \ln Z(\vec{\theta}|\mathcal{D}) \;=\; & -\frac{N}{2}\ln(2\pi) \\ & -\frac{1}{2}\sum_{i,j}(y(\mathbf{x}_i) - m(\mathbf{x}_i))\left[k(\mathbf{x}_i, \mathbf{x}_j)\right]^{-1}(y(\mathbf{x}_j) - m(\mathbf{x}_j)) \\ & -\frac{1}{2}\ln\left|\det\left[k(\mathbf{x}_i, \mathbf{x}_j)\right]\right| . \end{aligned} \tag{140}$$

The values of the hyperparameters can be fixed to their optimum values $\vec{\theta}_{\rm op}$, defined as those which maximise the hyperlikelihood:

$$\left.\frac{\partial Z(\vec{\theta}|\mathcal{D})}{\partial \theta^\mu}\right|_{\vec{\theta}=\vec{\theta}_{\rm op}} = 0\,. \tag{141}$$

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,

$$\Lambda_{\rm expanded}(\mathbf{x}, \vec{\theta}|\mathcal{D}) \;\propto \mathcal{L}(\mathbf{x}|\vec{\theta}, \mathcal{D}) Z(\vec{\theta}|\mathcal{D}). \tag{142}$$

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. As mentioned above, 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. (138). 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

$$\begin{bmatrix} y(\mathbf{x}_i) \\ y(\mathbf{z}_j) \end{bmatrix} \sim N\left(\boldsymbol{m}, \boldsymbol{\Sigma}\right) , \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{K} & \boldsymbol{K}_* \\ \boldsymbol{K}_*^{\mathrm{T}} & \boldsymbol{K}_{**} \end{pmatrix} , \tag{143}$$

where $\boldsymbol{K}$ is defined in Eq. (137) and the matrices $\boldsymbol{K}_*$ and $\boldsymbol{K}_{**}$ are defined as

$$[\boldsymbol{K}_*]_{ij} = k(\mathbf{x}_i, \mathbf{z}_j) , \quad [\boldsymbol{K}_{**}]_{ij} = k(\mathbf{z}_i, \mathbf{z}_j) . \tag{144}$$

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

$$p(\{y(\mathbf{z}_i)\}) \propto \exp\left[-\frac{1}{2}\sum_{j,k}(y(\mathbf{z}_j) - \mu_j)\Sigma^{-1}_{jk}(y(\mathbf{x}_k) - \mu_k)\right] \tag{145}$$

where the GPR mean and its associated error are given by

$$\mu_i = m(\mathbf{z}_i) + \sum_{j,k}[\boldsymbol{K}_*]_{ji}\left[\boldsymbol{K}^{-1}\right]_{jk}(\tilde{y}(\mathbf{x}_k) - m(\mathbf{x}_k)) , \tag{146}$$

$$\Sigma_{ij} = [\boldsymbol{K}_{**}]_{ij} - \sum_{k,l}[\boldsymbol{K}_*]_{ki}\left[\boldsymbol{K}^{-1}\right]_{kl}[\boldsymbol{K}_*]_{lj} . \tag{147}$$

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 requirement on a covariance function is that it is *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. However, there are other properties which are not required, but are still desirable.

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 $\tau \equiv |\vec{\tau}| = |\mathbf{x}_1 - \mathbf{x}_2|$.[1] Isotropy of a GP implies stationarity, but the converse is not true.

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.

[1] 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. (139), 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 $k(\mathbf{x}_1, \mathbf{x}_2) = \int \mathrm{d}\vec{u}\, k_{\vec{u}}(\mathbf{x}_1) k_{\vec{u}}(\mathbf{x}_2)$.[2] Applying this procedure to a D-dimensional SE function generates a non-stationary analogue

$$k(\mathbf{x}_i, \mathbf{x}_j) \;=\; \sigma_f \left|\mathcal{G}^i\right|^{1/4} \left|\mathcal{G}^j\right|^{1/4} \left|\frac{\mathcal{G}^i + \mathcal{G}^j}{2}\right|^{-1/2} \exp\left(-\frac{1}{2} Q_{ij}\right) , \tag{148}$$

where

$$Q_{ij} \;=\; (\mathbf{x}_i - \mathbf{x}_j)^a (\mathbf{x}_i - \mathbf{x}_j)^b \left(\frac{\mathcal{G}^i_{ab} + \mathcal{G}^j_{ab}}{2}\right)^{-1} , \tag{149}$$

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. (148) 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

$$g_{ab}(\mathbf{x}) = g_{ab}(\mathbf{x}_0) + (\mathbf{x}^c - \mathbf{x}^c_0) \left.\frac{\partial g_{ab}(\mathbf{x})}{\partial \lambda^c}\right|_{\mathbf{x}=\mathbf{x}_0} + \ldots \tag{150}$$

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.

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

[2] 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 $I = \int \mathrm{d}\vec{u} \sum_{i,j} a_i a_j k_{\vec{u}}(\mathbf{x}_i) k_{\vec{u}}(\mathbf{x}_j) = \int \mathrm{d}\vec{u} \left(\sum_i a_i k_{\vec{u}}(\mathbf{x}_i)\right)^2 \geq 0$.

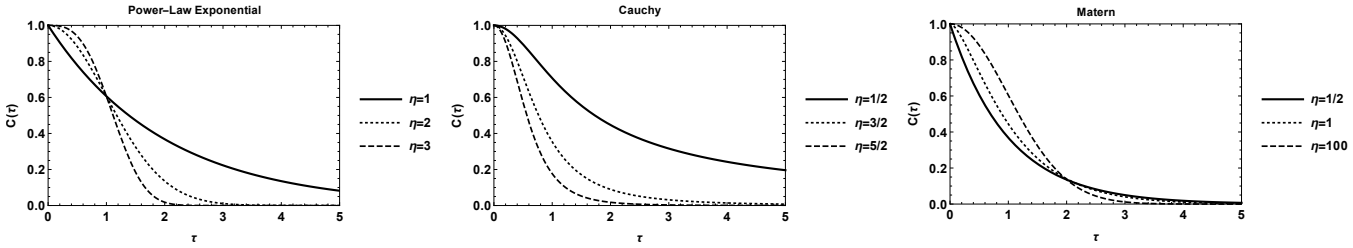


Figure 51: Plots of the different generalisations of the SE covariance function discussed in Sec. 11.2.2. The left-hand panel shows the PLE function, the centre panel shows the Cauchy function, and the right-hand panel shows the Matérn function; in all cases the value of σ_f was fixed to unity. In each 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. Figure reproduced from Moore et al. (2015).

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, including

- The **squared-exponential** covariance function (which has already been introduced), given by
$$k_{\rm SE}(\tau) = \sigma_f^2 \exp\left(-\frac{1}{2}\tau^2\right) . \tag{151}$$

- The **power-law exponential** (PLE) covariance function, given by
$$k_{\rm PLE}(\tau) = \sigma_f^2 \exp\left(-\frac{1}{2}\tau^\eta\right) , \tag{152}$$
where $0 < \eta \leq 2$. The PLE reduces to the SE in the case $\eta = 2$.

- The **Cauchy** covariance function, given by
$$k_{\rm Cauchy}(\tau) = \frac{\sigma_f^2}{(1 + \tau^2/2\eta)^\eta} , \tag{153}$$
where $\eta > 0$. This recovers the SE function in the limit $\eta \to \infty$.

- The **Matérn** covariance function, given by
$$k_{\rm Mat}(\tau) = \frac{\sigma_f^2 2^{1-\eta}}{\Gamma(\eta)} \left(\sqrt{2\eta}\,\tau\right)^\eta K_\eta \left(\sqrt{2\eta}\,\tau\right) , \tag{154}$$
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. 51 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, due to its differentiability properties.

We will see in section 11.3 that the mean-square differentiability of a GP is determined by the differentiability of its covariance function at $\tau = 0$. The SE covariance function is infinitely differentiable at $\tau = 0$, and so the corresponding GP is infinitely (mean-square) 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 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. (154) is ζ-times mean-square 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 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;

$$k(\tau) > 0 \quad \forall \tau \in [0, \infty) \,. \tag{155}$$

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 can still become prohibitive for large training sets. For large training sets the determinant of the covariance matrix is also typically small which contributes to making the covariance matrix hard to invert.

One potential way around these issues is to consider a covariance function with compact support,

$$\begin{aligned} k(\tau) &> 0 \quad \tau \in [0, T] \,, \\ k(\tau) &= 0 \quad \forall \tau \in (T, \infty) \,, \end{aligned} \tag{156}$$

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.

It is possible to construct covariance functions which have the requisite properties and satisfy the compact support condition in Eq. (156). These are typically based on polynomials. One such series of polynomials was proposed by Wendland. These have the property that they are positive definite in $\mathbb{R}^D$ and are $2q$-times differentiable at the origin. Therefore the discrete parameter q is in some sense analogous to the η hyperparameter of the Matérn

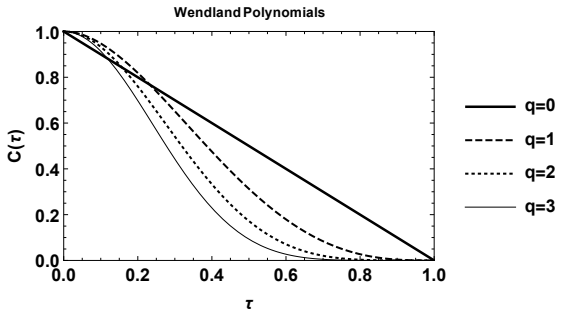


Figure 52: Plots of the first few Wendland polynomial covariance functions. All these functions have compact support, $k(\tau) = 0$ for $\tau > 1$. As the value of q increases the functions become smoother near the origin. Figure reproduced from Moore et al. (2015).

covariance function in that it controls the smoothness of the GP. Defining β to be

$$\beta = \left\lfloor \frac{D}{2} \right\rfloor + q + 1 \tag{157}$$

and where $\Theta(x)$ denotes the Heaviside step function, the first few Wendland polynomials $k_{D,q}(\tau)$ are given by,

$$\begin{aligned}
k_{D,0}(\tau) &= \sigma_f^2 \Theta(1-\tau)(1-\tau)^{\beta}\,, && (158)\\
k_{D,1}(\tau) &= \sigma_f^2 \Theta(1-\tau)(1-\tau)^{\beta+1}\left[1+(\beta+1)\,\tau\right]\,, && (159)\\
k_{D,2}(\tau) &= \frac{\sigma_f^2}{3}\Theta(1-\tau)(1-\tau)^{\beta+2}\left[\,3+(3\beta+6)\,\tau\right.\\
&\quad \left.+\left(\beta^2+4\beta+3\right)\tau^2\right]\,, && (160)\\
k_{D,3}(\tau) &= \frac{\sigma_f^2}{15}\Theta(1-\tau)(1-\tau)^{\beta+3}\left[15+(15\beta+45)\,\tau\right.\\
&\quad +\left(6\beta^2+36\beta+45\right)\tau^2\\
&\quad \left.+\left(\beta^3+9\beta^2+23\beta+15\right)\tau^3\right]\,. && (161)
\end{aligned}$$

The Wendland polynomials are plotted in Fig. 52. Other types of covariance functions with compact support have also been proposed and explored in the literature, but we will not discuss 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 *mean-square* (MS) continuous at $\mathbf{x}_*$ if

$$\lim_{\ell\to\infty} \mathbb{E}\left[(Y(\mathbf{x}_\ell) - Y(\mathbf{x}_*)|Y(\mathbf{x}_\ell) - Y(\mathbf{x}_*))\right] = 0\,, \tag{162}$$

where $\mathbb{E}[\ldots]$ denotes the expectation of the enclosed quantity over realisations of the GP. Here we are using $(a|b)$ to denote the inner product on the output space of the GP. Normally this will be a vector of real or complex values, in which case this reduces to the usual norm.

MS continuity implies continuity in the mean,

$$\lim_{\ell\to\infty} \mathbb{E}\left[Y(\mathbf{x}_\ell) - Y(\mathbf{x}_*)\right] = 0\,. \tag{163}$$

There are other notions of continuity of GPs used in the literature, but the notion of MS continuity relates most easily to the properties of the covariance function.

The mean and the covariance of a GP are defined as

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[Y(\mathbf{x})]\,, \\ k(\mathbf{x}_1,\mathbf{x}_2) &= \mathbb{E}[(Y(\mathbf{x}_1) - m(\mathbf{x}_1)|Y(\mathbf{x}_2) - m(\mathbf{x}_2))]\,. \end{aligned} \tag{164}$$

Using these, Eq. (162) can be written as

$$\begin{aligned} \lim_{\ell\to\infty} \{&k(\mathbf{x}_*,\mathbf{x}_*) - 2k(\mathbf{x}_\ell,\mathbf{x}_*) + k(\mathbf{x}_\ell,\mathbf{x}_\ell) \\ &+ (m(\mathbf{x}_*) - m(\mathbf{x}_\ell)|m(\mathbf{x}_*) - m(\mathbf{x}_\ell))\} \;=\; 0\,, \end{aligned} \tag{165}$$

and using the continuity of the mean in Eq. (163) gives

$$\lim_{\ell\to\infty} \left[k(\mathbf{x}_*,\mathbf{x}_*) - 2k(\mathbf{x}_\ell,\mathbf{x}_*) + k(\mathbf{x}_\ell,\mathbf{x}_\ell)\right] = 0\,. \tag{166}$$

This condition is satisfied if the covariance function, $k(\mathbf{x}_1,\mathbf{x}_2)$, 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.[3] 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. (162), the notion of taking the MS derivative of a GP is defined as

$$\frac{\partial Y(\mathbf{x})}{\partial \mathbf{x}^a} = \operatorname*{l.i.m}_{\epsilon\to 0} X_a(\mathbf{x},\epsilon)\,, \tag{167}$$

where l.i.m is read limit in MS and

$$X_a(\mathbf{x},\epsilon) = \frac{Y(\mathbf{x} + \epsilon\,\hat{e}_a) - Y(\mathbf{x})}{\epsilon} \tag{168}$$

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

$$\begin{aligned} K_\epsilon(\mathbf{x}_1,\mathbf{x}_2) &= \mathbb{E}\left[(X_a(\mathbf{x}_1,\epsilon) - \Xi(\mathbf{x}_1,\epsilon)|\right. \\ &\qquad \left. X_a(\mathbf{x}_2,\epsilon) - \Xi(\mathbf{x}_2,\epsilon))\right] \end{aligned} \tag{169}$$

where $\Xi_a(\mathbf{x},\epsilon) = \mathbb{E}[X_a(\mathbf{x},\epsilon)]$. It then follows that

$$\begin{aligned} K_\epsilon(\mathbf{x}_1,\mathbf{x}_2) &= \frac{k(\mathbf{x}_1+\epsilon,\mathbf{x}_2+\epsilon) - k(\mathbf{x}_1,\mathbf{x}_2+\epsilon)}{\epsilon^2} \\ &\quad + \frac{k(\mathbf{x}_1+\epsilon,\mathbf{x}_2) - k(\mathbf{x}_1,\mathbf{x}_2)}{\epsilon^2}\,. \end{aligned} \tag{170}$$

[3] 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. (167), 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

$$\frac{\partial^\zeta Y(\mathbf{x})}{\partial \mathbf{x}^{a_1} \partial \mathbf{x}^{a_2} \ldots \partial \mathbf{x}^{a_\zeta}}\,, \tag{171}$$

is given by the 2ζ-times differentiated covariance function

$$\frac{\partial^{2\zeta} k(\mathbf{x}_1, \mathbf{x}_2)}{\partial \mathbf{x}_1^{a_1} \partial \mathbf{x}_2^{a_1} \partial \mathbf{x}_1^{a_2} \partial \mathbf{x}_2^{a_2} \ldots \partial \mathbf{x}_1^{a_\zeta} \partial \mathbf{x}_2^{a_\zeta}}\,. \tag{172}$$

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.[4]

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 53. 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 54 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 55 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 $\sim$ 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

[4] 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.

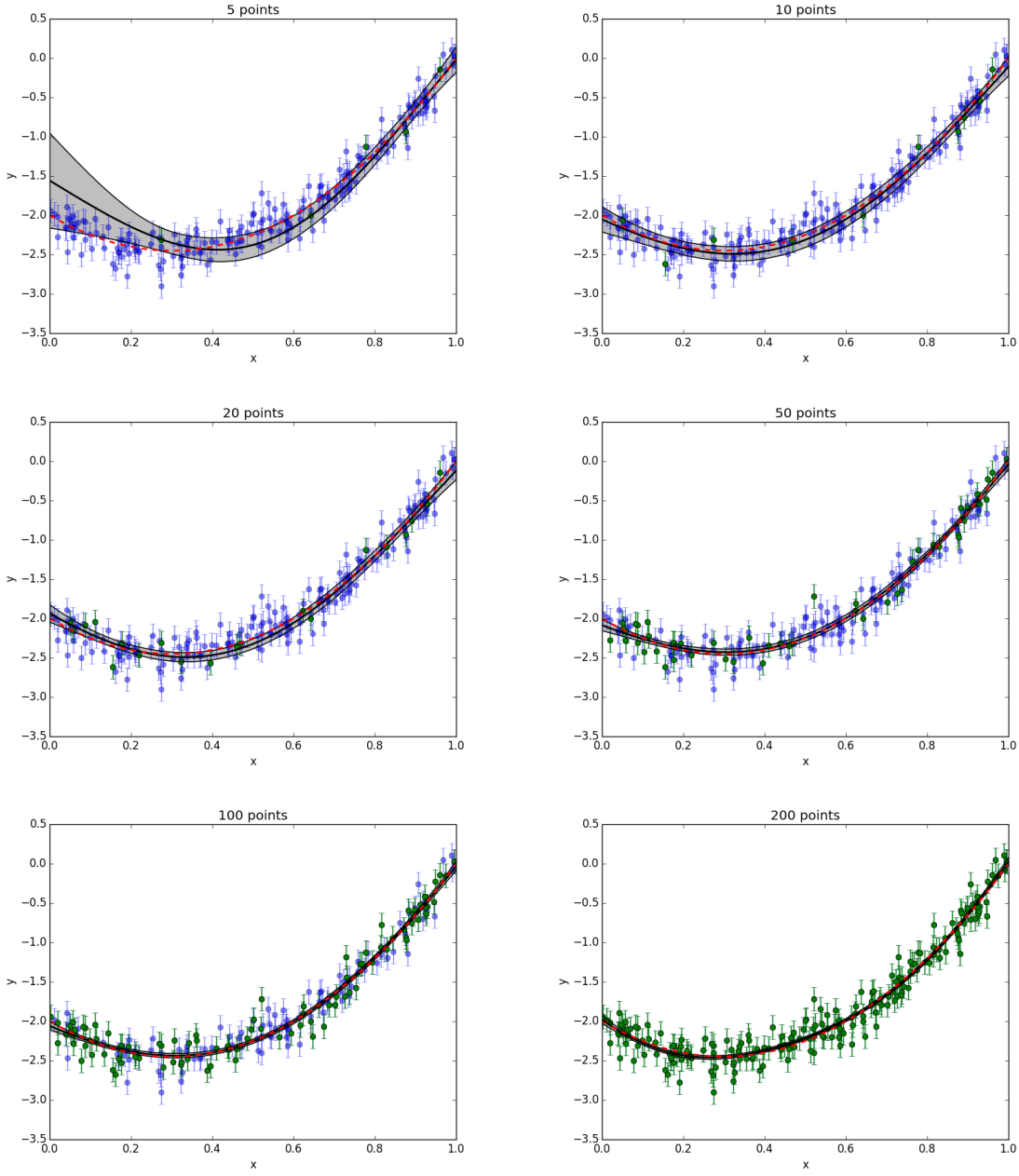


Figure 53: Gaussian process fit to noisy measurements of a quadratic, for different sizes of training set, as stated in the title of each panel.

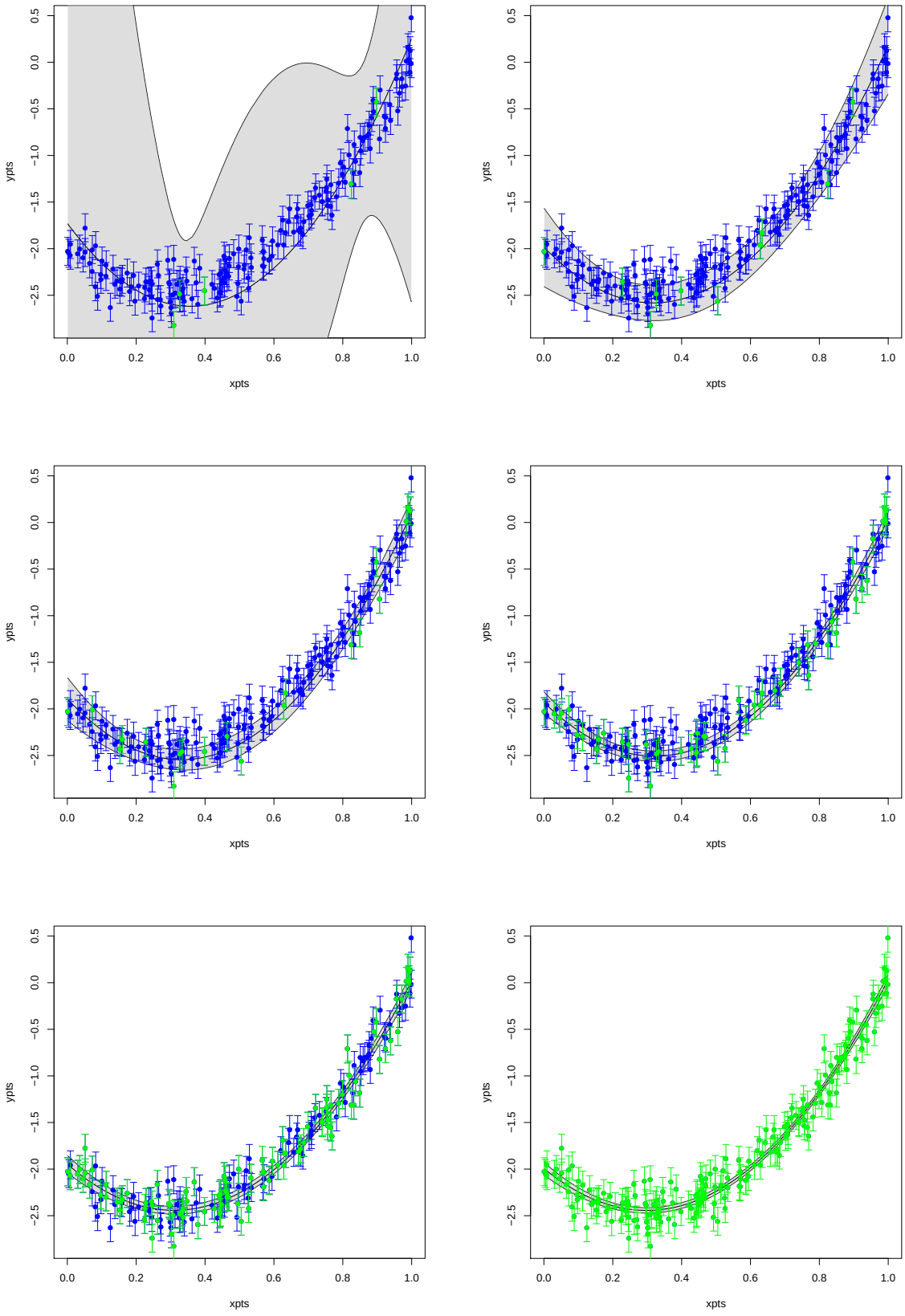


Figure 54: As Figure 53, but now fitting a quadratic linear model to the same data.

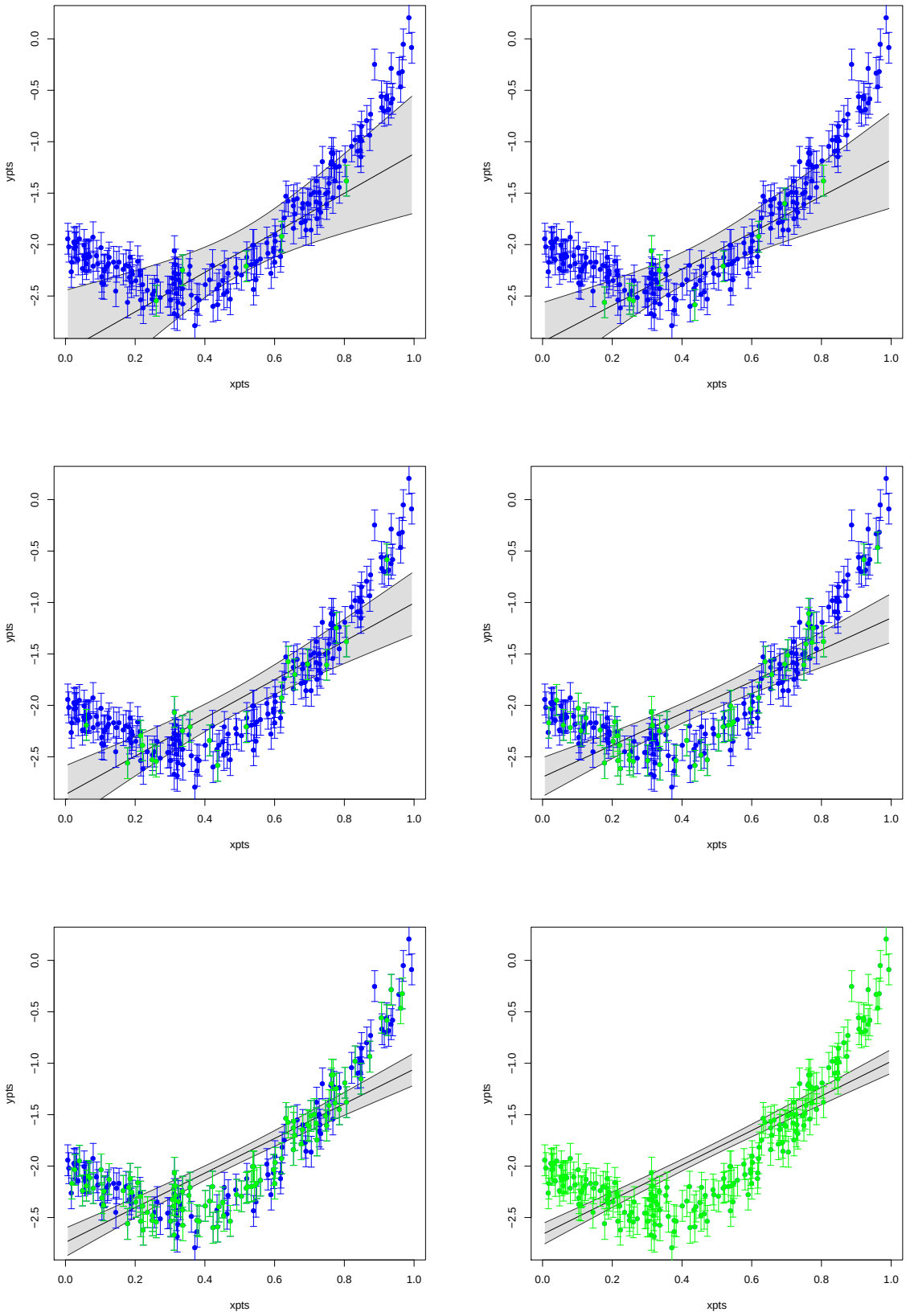


Figure 55: As Figure 53, but now fitting a linear model to the same data.

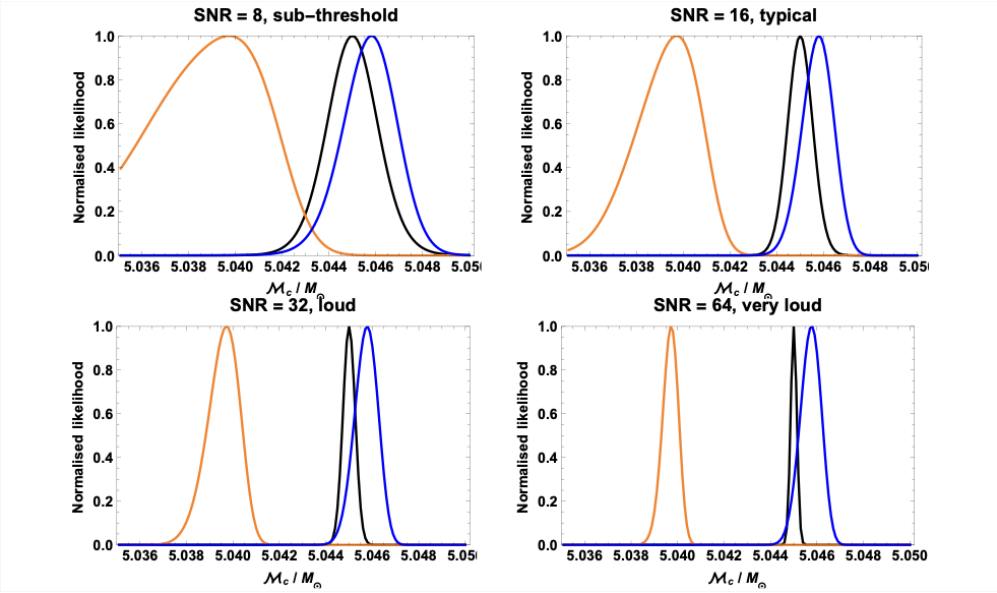


Figure 56: Comparison between uncorrected, corrected and "true" likelihood for inference with waveform models that include model error. The corrected likelihood uses a Gaussian process to model the waveform error and then marginalises this out of the likelihood. Reproduced from Moore et al. (2015).

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

$$\mathcal{L}(\vec{\lambda}) \propto \frac{1}{\sqrt{1+\sigma^2(\vec{\lambda})}} \exp\left(-\frac{1}{2}\frac{\left\|s - H(\vec{\lambda}) + \mu(\vec{\lambda})\right\|^2}{1+\sigma^2(\vec{\lambda})}\right). \tag{173}$$

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 56.

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 57 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

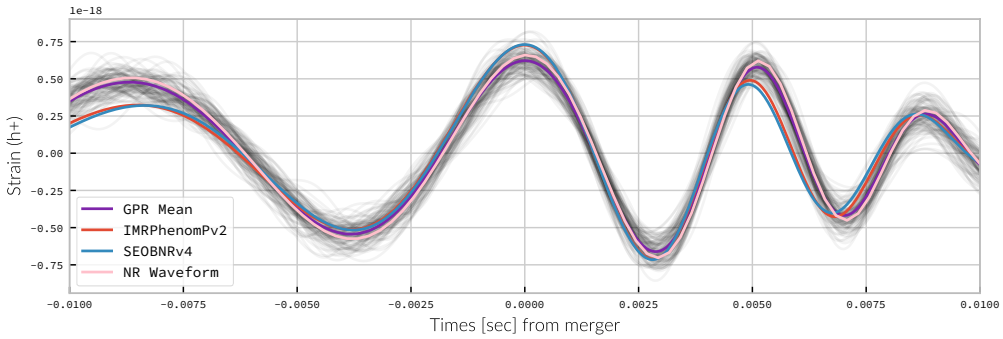


Figure 57: Comparison of several approximate waveform models to random draws from a Gaussian process interpolant trained on numerical relativity simulations. Reproduced from Williams et al. (2020).

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 58 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 59 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 $X \sim \mathrm{DP}(P, a)$ 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

$$(X(B_1), X(B_2), \ldots, X(B_n)) \sim \mathrm{Dir}(aP(B_1), aP(B_2), \ldots, aP(B_n)). \tag{174}$$

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

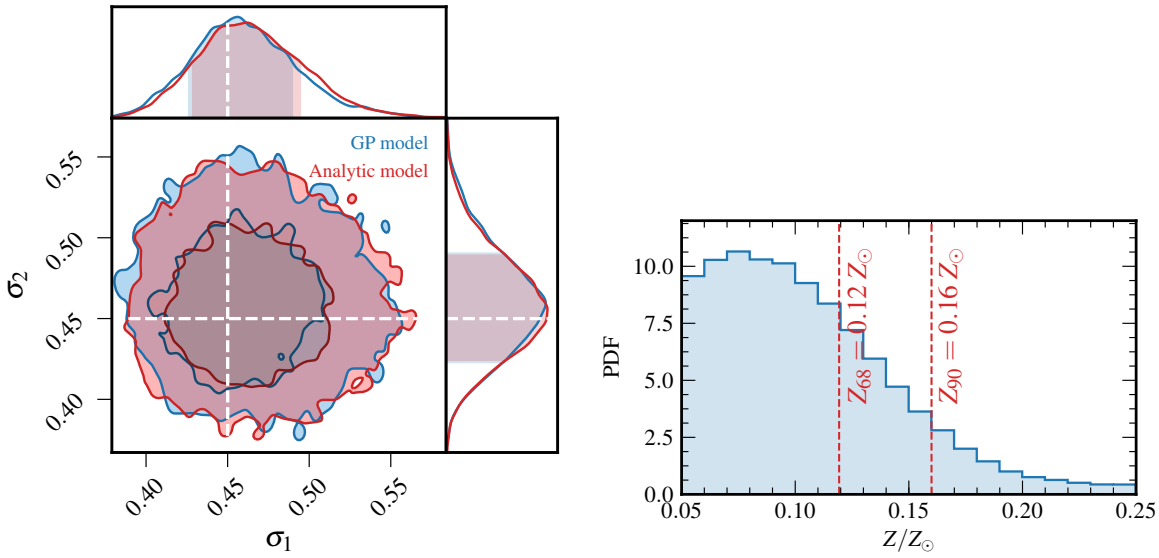


Figure 58: Posteriors on physical parameters of the astrophysical source population inferred form simulated observations of binaries. Inference relied on a Gaussian process model that interpolated the output of the population synthesis codes over the astrophysical parameter space. Reproduced from Taylor & 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 60 and 61 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 60 and 61 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

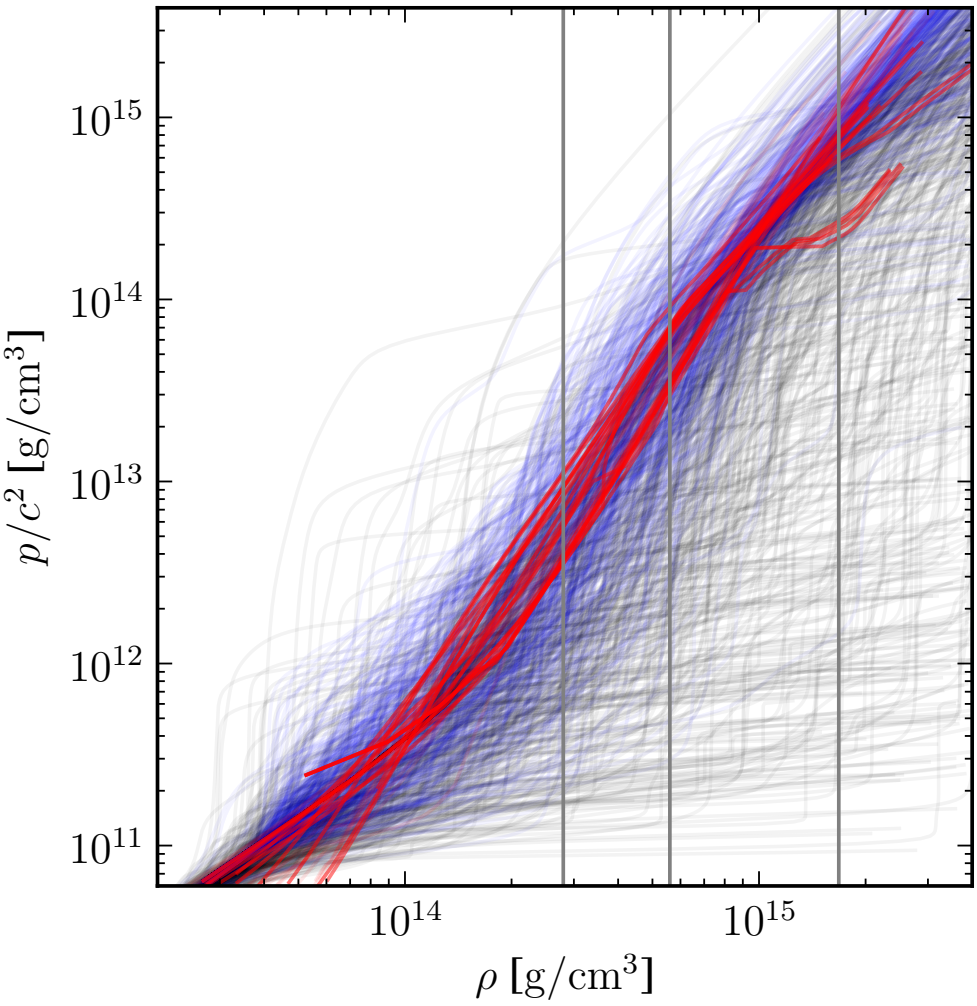


Figure 59: Random draws from a Gaussian process model of the equation of state of a neutron star. Reproduced from Essick et al. (2019).

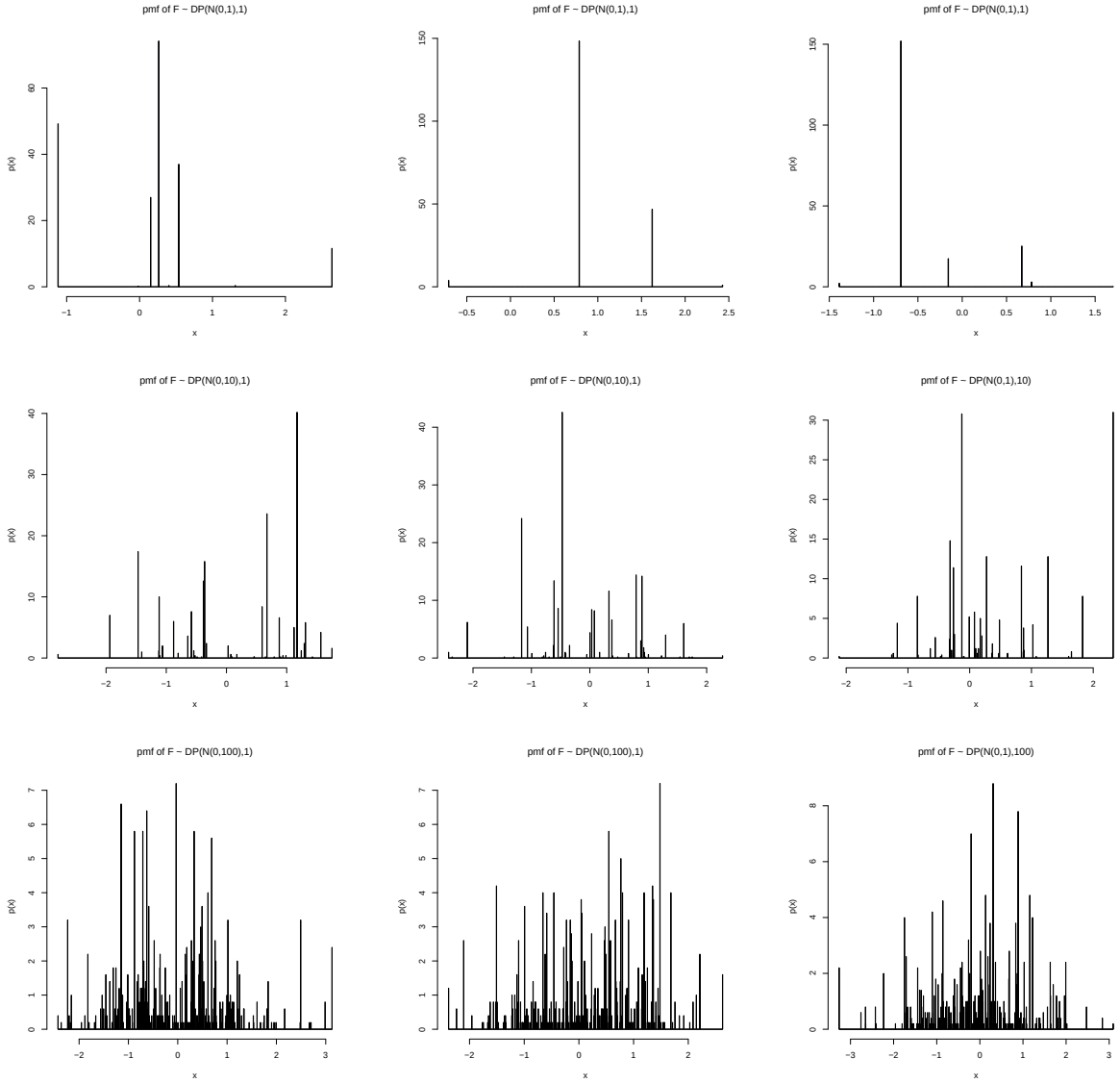


Figure 60: Sample realisations of a Dirichlet process, $X \sim \text{DP}(N(0,1), a)$, for $a = 1$ (top row), $a = 10$ (middle row) and $a = 100$ (bottom row). In each figure we show 1000 samples from the given realisation of the Dirichlet process. Within each row, the figures show three distinct realisations of the stated Dirichlet process.

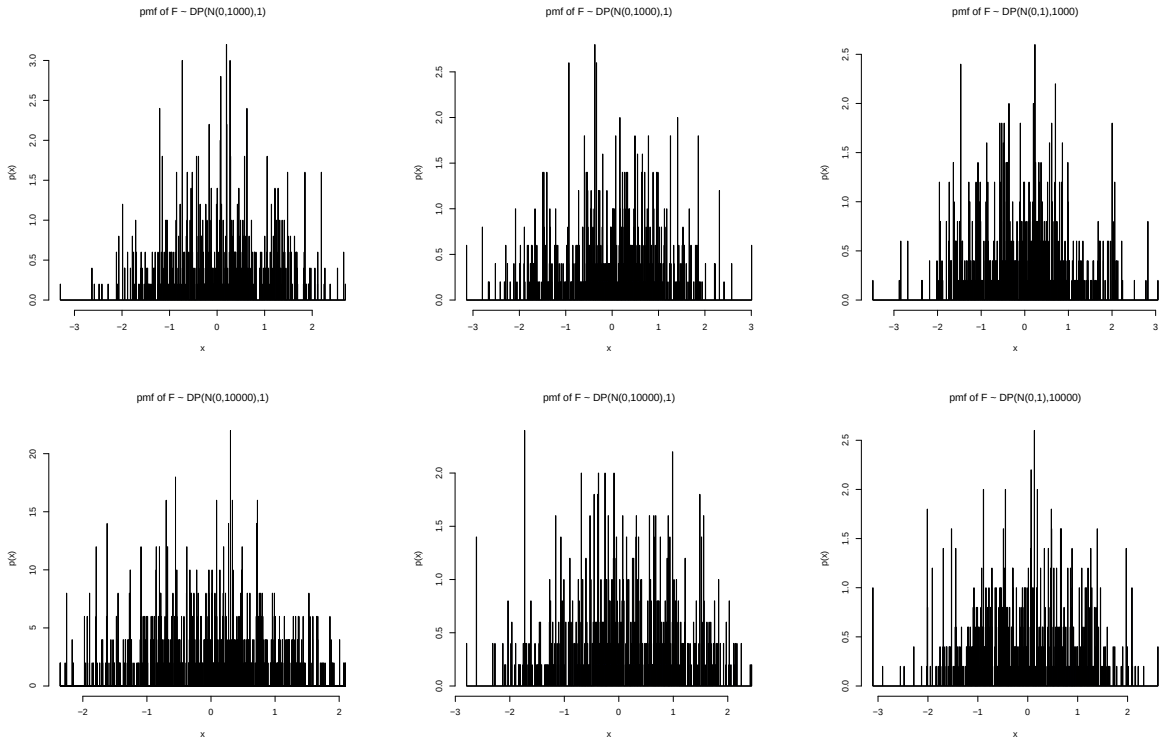


Figure 61: As in Figure 60, these figures show sample realisations of a Dirichlet process, $X \sim \text{DP}(N(0,1), a)$, for $a = 1000$ (top row) and $a = 10000$ (bottom row). In each figure we show 1000 samples from the given realisation of the Dirichlet process. Within each row, the figures show three 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

$$
\begin{aligned}
X &= \left(\sum_{l=1}^{L_H} p_l \delta_{U_l}\right) + \left(1 - \sum_{l=1}^{L_H} p_l\right) \delta_{U_0} \\
p_1 &= V_1, \qquad p_l = \left(\prod_{j=1}^{l-1}(1-V_j)\right) V_l, \;\; l \geq 2, \qquad p_0 = 1 - \sum_{l=1}^{L_H} p_l \\
V_l &\sim \text{Beta}(1,a), \;\; l = 1,\ldots,L_H, \qquad U_l \sim P, \;\; l = 0,1,\ldots,L_H,
\end{aligned} \tag{175}
$$

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

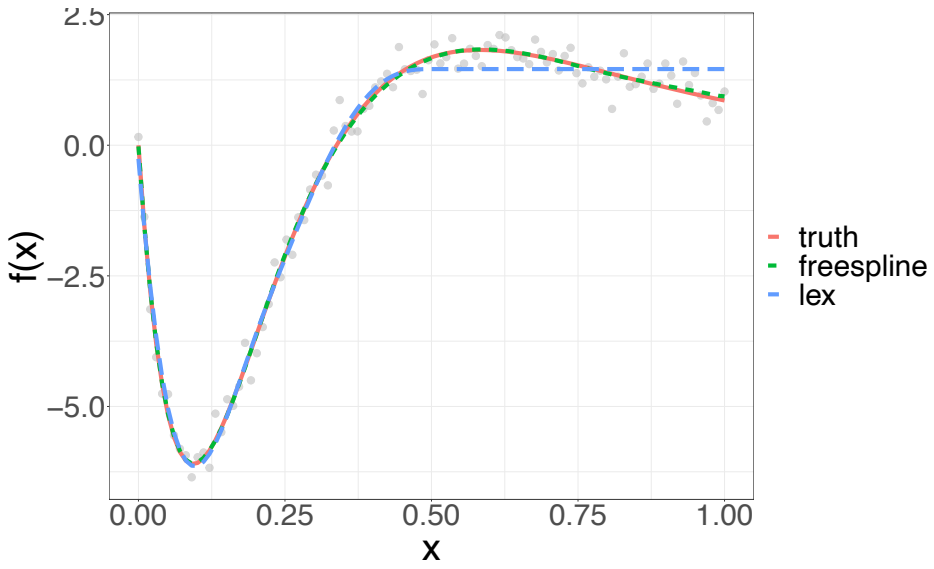


Figure 62: Nonparametric regression fit to noisy measurements of the function $f(x) = 26\exp(-3.25x) - 4\exp(-6.5x) + 3\exp(-9.75x)$ using the freespline algorithm with a Dirichlet process prior on the probability density determining the knot locations. Figure reproduced from Edwards & Gair (2020).

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 62 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 63 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

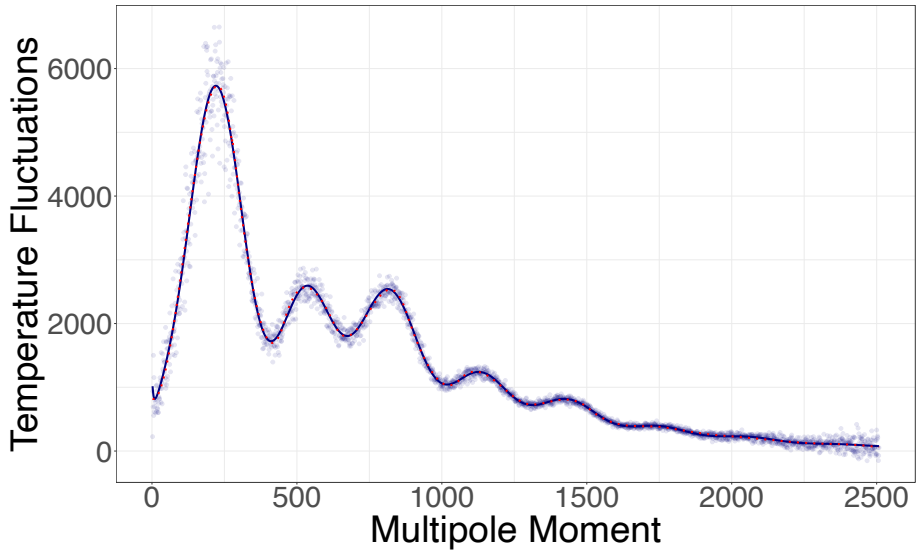


Figure 63: Nonparametric regression fit to the CMB temperature power spectrum, as measured by Planck. The dashed red line is the freespline fit to the data, while the blue line is the prediction of the best fit cosmological model. Figure reproduced from Edwards & Gair (2020).

previously generated by LALInference. In Figure 64 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.

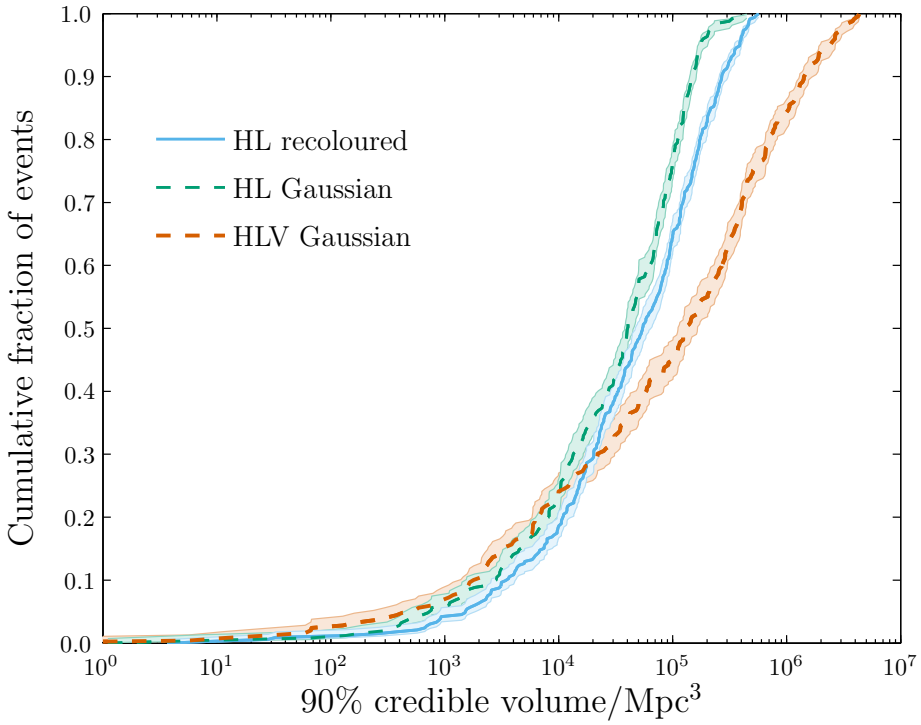


Figure 64: Cumulative distribution function of 90% credible volumes for events observed by the ground-based detector network. The credible volumes were computed by fitting a Dirichlet Process Gaussian Mixture Model to posterior samples generated by LALInference. Figure reproduced from Del Pozzo et al. (2018).