Introduction to Nonparametric Analysis in Time Series Econometrics

Yongmiao Hong

2020

This is Chapter 6 of a manuscript entitled as *Modern Time Series Analysis: Theory and Applications* written by the author. We will introduce some popular nonparametric methods, particularly the kernel smoothing method and the local polynomial smoothing method, to estimate functions of interest in time series contexts, such as probability density functions, autoregression functions, spectral density functions, and generalized spectral density functions. Empirical applications of these functions crucially depend on the consistent estimation of these functions. We will discuss the large sample statistical properties of nonparametric estimators in various contexts.

Key words: Asymptotic normality, bias, boundary problem, consistency, curse of dimensionality, density function, generalized spectral density, global smoothing, integrated mean squared error, law of large numbers, local polynomial smoothing, local smoothing, locally stationary time series model, mean squared error, kernel method, regression function, series approximation, smoothing, spectral density function, Taylor series expansion, variance.

Reading Materials and References

This lecture note is self-contained. However, the following references will be useful for learning nonparametric analysis.

(1) Nonparametric Analysis in Time Domain

- Silverman, B. (1986): *Nonparametric Density Estimation and Data Analysis.* Chapman and Hall: London.
- Härdle, W. (1990): *Applied Nonparametric Regression.* Cambridge University Press: Cambridge.
- Fan, J. and Q. Yao (2003), *Nonlinear Time Series: Parametric and Nonparametric Methods*, Springer: New York.

(2) Nonparametric Methods in Frequency Domain

- Priestley, M. (1981), *Spectral Analysis and Time Series.* Academic Press: New York.
- Hannan, E. (1970), *Multiple Time Series*, John Wiley: New York.

1 Motivation

Suppose $\{X_t\}$ is a strictly stationary process with marginal probability density function $g(x)$ and pairwise joint probability density function $f_j(x, y)$, and a random sample $\{X_t\}_{t=1}^T$ of size T is observed. Then,

- How to estimate the marginal pdf $g(x)$ of $\{X_t\}$?
- How to estimate the pairwise joint pdf $f_j(x, y)$ of (X_t, X_{t-j})?
- How to estimate the autoregression function $r_j(x) = E(X_t|X_{t-j} = x)$?
- How to estimate the spectral density $h(\omega)$ of $\{X_t\}$?
- How to estimate the generalized spectral density $f(\omega, u, v)$ of $\{X_t\}$?
- How to estimate the bispectral density $b(\omega_1, \omega_2)$?
- How to estimate a nonlinear autoregressive conditional heteroskedastic model

$$X_t = \mu(X_{t-1}, ..., X_{t-p}) + \sigma(X_{t-1}, ..., X_{t-q})\varepsilon_t, \qquad \{\varepsilon_t\} \sim i.i.d.(0, 1),$$

  where $\mu(\cdot)$ and $\sigma(\cdot)$ are unknown functions of the past information. Under certain regularity conditions, $\mu(\cdot)$ is the conditional mean of X_t given $I_{t-1} = \{X_{t-1}, X_{t-2}, ...\}$ and $\sigma^2(\cdot)$ is the conditional variance of X_t given I_{t-1}.

- How to estimate a semi-nonparametric functional coefficient autoregressive process

$$X_t = \sum_{j=1}^{p} \alpha_j(X_{t-d})X_{t-j} + \varepsilon_t, \qquad E(\varepsilon_t|I_{t-1}) = 0 \text{ a.s.},$$

  where $\alpha_j(\cdot)$ is unknown, and $d > 0$ is a time lag parameter?

- How to estimate a nonparametric additive autoregressive process

$$X_t = \sum_{j=1}^{p} \mu_j(X_{t-j}) + \varepsilon_t, \qquad E(\varepsilon_t|I_{t-1}) = 0 \text{ a.s.},$$

  where the $\mu_j(\cdot)$ functions are unknown?

- How to estimate a locally linear time-varying regression model

$$Y_t = X_t'\beta(t/T) + \varepsilon_t,$$

  where $\beta(\cdot)$ is an unknown smooth deterministic function of time?

- How to use these estimators in economic and financial applications?

Nonparametric estimation is often called **nonparametric smoothing**, since a key parameter called smoothing parameter is used to control the degree of the estimated curve. Nonparametric smoothing first arose from spectral density estimation in time series analysis. In a discussion of the seminal paper by Bartlett (1946), Henry Daniels suggested that a possible improvement on spectral density estimation could be made by smoothing the periodogram (see Chapter 3), which is the squared discrete Fourier transform of the random sample $\{X_t\}_{t=1}^T$. The theory and techniques were then systematically developed by Bartlett (1948,1950). Thus, smoothing techniques were already prominently featured in time series analysis more than 70 years ago.

In the earlier stage of nonlinear time series analysis (see Tong (1990)), the focus was on various nonlinear parametric forms, such as threshold autoregressive models, smooth transition autoregressive models, and Regime-switch Markov chain autoregressive models (see Chapter 8 for details). Recent interest has been mainly in nonparametric curve estimation, which does not require the knowledge of the functional form beyond certain smoothness conditions on the underlying function of interest.

Question: Why is nonparametric smoothing popular in statistics and econometrics?

There are several reasons for the popularity of nonparametric analysis. In particular, three main reasons are:

- Demands for nonlinear approaches;
- Availability of large data sets;
- Advance in computer technology.

Indeed, as Granger (1999) points out, the speed in computing technology increases much faster than the speed at which data grows.

To obtain basic ideas about nonparametric smoothing methods, we now consider two examples, one is the estimation of a regression function, and the other is the estimation of a probability density function.

Example 1 [Regression Function]: Consider the first order autoregression function

$$r_1(x) = E(X_t|X_{t-1} = x).$$

We can write

$$X_t = r_1(X_{t-1}) + \varepsilon_t,$$

where $E(\varepsilon_t|X_{t-1}) = 0$ by construction. We assume $E(X_t^2) < \infty$.

Suppose a sequence of bases $\{\psi_j(x)\}$ constitutes a complete orthonormal basis for the space of square-integrable functions. Then we can always decompose the function

$$r_1(x) = \sum_{j=0}^{\infty} \alpha_j \psi_j(x),$$

where the Fourier coefficient

$$\alpha_j = \int_{-\infty}^{\infty} r_1(x)\psi_j(x)dx,$$

which is the projection of $r_1(x)$ on the base $\psi_j(x)$.

Suppose there is a quadratic function $r_1(x) = x^2$ for $x \in [-\pi, \pi]$. Then

$$\begin{aligned} r_1(x) &= \frac{\pi^2}{3} - 4\left(\cos(x) - \frac{\cos(2x)}{2^2} + \frac{\cos(3x)}{3^2} - \cdots\right) \\ &= \frac{\pi^2}{3} - 4\sum_{j=1}^{\infty}(-1)^{j-1}\frac{\cos(jx)}{j^2}. \end{aligned}$$

For another example, suppose the regression function is a step function, namely

$$r_1(x) = \begin{cases} -1 & \text{if } -\pi < x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } 0 < x < \pi. \end{cases}$$

Then we can still expand it as an infinite sum of periodic series,

$$\begin{aligned} r_1(x) &= \frac{4}{\pi}\left[\sin(x) + \frac{\sin(3x)}{3} + \frac{\sin(5x)}{5} + \cdots\right] \\ &= \frac{4}{\pi}\sum_{j=0}^{\infty}\frac{\sin[(2j+1)x]}{(2j+1)}. \end{aligned}$$

In general, we do not assume that the function form of $r_1(x)$ is known, except that we still maintain the assumption that $r_1(x)$ is a square-integrable function. Because $r_1(x)$ is square-integrable, we have

$$\begin{aligned}\int_{-\infty}^{\infty} r_1^2(x)dx &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\alpha_j\alpha_k\int_{-\infty}^{\infty}\psi_j(x)\psi_k(x)dx \\ &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\alpha_j\alpha_k\delta_{j,k} \text{ by orthonormality} \\ &= \sum_{j=0}^{\infty}\alpha_j^2 < \infty,\end{aligned}$$

where $\delta_{j,k}$ is the Kronecker delta function: $\delta_{j,k} = 1$ if $j = k$ and 0 otherwise.

The squares summability implies $\alpha_j \to 0$ as $j \to \infty$, that is, α_j becomes less important as the order $j \to \infty$. This suggests that a truncated sum

$$r_{1p}(x) = \sum_{j=0}^{p}\alpha_j\psi_j(x)$$

can be used to approximate $r_1(x)$ arbitrarily well if p is sufficiently large. The approximation error, or the bias,

$$\begin{aligned}b_p(x) &\equiv r_1(x) - r_{1p}(x) \\ &= \sum_{j=p+1}^{\infty}\alpha_j\psi_j(x) \\ &\to 0\end{aligned}$$

as $p \to \infty$.

However, the coefficient α_j is unknown. To obtain a feasible estimator for $r_1(x)$, we consider the following sequence of truncated regression models

$$X_t = \sum_{j=0}^{p}\beta_j\psi_j(X_{t-1}) + \varepsilon_{pt},$$

where $p \equiv p(T) \to \infty$ is the number of series terms that depends on the sample size T. We need $p/T \to 0$ as $T \to \infty$, i.e., the number of p is much smaller than the sample size T. Note that the regression error ε_{pt} is not the same as the true innovation ε_t for each given p. Instead, it contains the true innovation ε_t and the bias $b_p(X_{t-1})$.

The ordinary least squares estimator

$$\begin{aligned}\hat{\beta} &= (\Psi'\Psi)^{-1}\Psi'X \\ &= \left(\sum_{t=2}^{T}\psi_t\psi_t'\right)^{-1}\sum_{t=2}^{T}\psi_t X_t,\end{aligned}$$

where

$$\Psi = (\psi_1', ..., \psi_T')'$$

is a $T \times p$ matrix, and

$$\psi_t = [\psi_0(X_{t-1}), \psi_1(X_{t-1}), ..., \psi_p(X_{t-1})]'$$

is a $p \times 1$ vector. The series-based regression estimator is

$$\hat{r}_{1p}(x) = \sum_{j=0}^{p}\hat{\beta}_j\psi_j(x).$$

To ensure that $\hat{r}_{1p}(x)$ is asymptotically unbiased, we must let $p = p(T) \to \infty$ as $T \to \infty$ (e.g., $p = \sqrt{T}$). However, if p is too large, the number of estimated parameters will be too large, and as a consequence, the sampling variation of $\hat{\beta}$ will be large (i.e., the estimator $\hat{\beta}$ is imprecise.) We must choose an appropriate $p = P(T)$ so as to balance the bias and the sampling variation. The truncation order p is called a smoothing parameter because it controls the smoothness of the estimated function $\hat{r}_{1p}(x)$. In general, for any given sample, a large p will give a smooth estimated curve whereas a small p will give a wiggly estimated curve. If p is too large such that the variance of $\hat{r}_{1p}(x)$ is larger than its squared bias, we call that there exists oversmoothing. In contrast, if p is too sall such that the variance of $\hat{r}_{1p}(x)$ is smaller than its squared bias, then we call that there exists undersmoothing. Optimal smoothing is achieved when the variance of $\hat{r}_{1p}(x)$ balances its squared bias. The series estimator $\hat{r}_{1p}(x)$ is called a global smoothing method, because once p is given, the estimated function $\hat{r}_{1p}(x)$ is determined over the entire domain of X_t.

Under suitable regularity conditions, $\hat{r}_{1p}(x)$ will consistently estimate the unknown function $r_1(x)$ as the sample size T increases. This is called nonparametric estimation because no parametric functional form is imposed on $r_1(x)$.

The base functions $\{\psi_j(\cdot)\}$ can be the Fourier series (i.e., the sin and cosine functions), and B-spline functions if X_t has a bounded support. See (e.g.) Andrews (1991, *Econometrica*) and Hong and White (1995, *Econometrica*) for applications.

Example 2 [Probability Density Function]: Suppose the PDF $g(x)$ of X_t is a smooth function with unbounded support. We can expand

$$g(x) = \phi(x) \sum_{j=0}^{\infty} \beta_j H_j(x),$$

where the function

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$$

is the $N(0,1)$ density function, and $\{H_j(x)\}$ is the sequence of Hermite polynomials, defined as

$$(-1)^j \frac{d^j}{dx^j}\Phi(x) = -H_{j-1}(x)\phi(x) \text{ for } j > 0,$$

where $\Phi(\cdot)$ is the $N(0,1)$ CDF. For example,

$$\begin{aligned} H_0(x) &= 1, \\ H_1(x) &= x, \\ H_2(x) &= (x^2 - 1) \\ H_3(x) &= x(x^2 - 3), \\ H_4(x) &= x^4 - 6x^2 + 3. \end{aligned}$$

See, for example, Magnus, Oberhettinger and Soni (1966, Section 5.6) and Abramowitz and Stegun (1972, Ch.22).

Here, the Fourier coefficient

$$\beta_j = \int_{-\infty}^{\infty} g(x) H_j(x) \phi(x) dx.$$

Again, $\beta_j \to 0$ as $j \to \infty$ given $\sum_{j=0}^{\infty} \beta_j^2 < \infty$.

The $N(0,1)$ PDF $\phi(x)$ is the leading term to approximate the unknown density $g(x)$, and the Hermite polynomial series will capture departures from normality (e.g., skewness and heavy tails).

To estimate $g(x)$, we can consider the sequence of truncated probability densities

$$g_p(x) = C_p^{-1} \phi(x) \sum_{j=0}^{p} \beta_j H_j(x),$$

where the constant

$$C_p = \sum_{j=0}^{p} \beta_j \int H_j(x) \phi(x) dx$$

is a normalization factor to ensure that $g_p(x)$ is a PDF for each p. The unknown parameters $\{\beta_j\}$ can be estimated from the sample $\{X_t\}_{t=1}^T$ via the maximum likelihood estimation (MLE) method. For example, suppose $\{X_t\}$ is an IID sample. Then

$$\hat{\beta} = \arg\max_{\beta} \sum_{t=1}^{T} \ln \hat{g}_p(X_t)$$

To ensure that

$$\hat{g}_p(x) = \hat{C}_p^{-1}\phi(x)\sum_{j=0}^{p}\hat{\beta}_j H_j(x)$$

is asymptotically unbiased, we must let $p = p(T) \to \infty$ as $T \to \infty$. However, p must grow more slowly than the sample size T grows to infinity so that the sampling variation of $\hat{\beta}$ will not be too large.

For the use of Hermite Polynomial series expansions, see (e.g.) Gallant and Tauchen (1996, *Econometric Theory*), Aït-Sahalia (2002, *Econometrica*), and Cui, Hong and Li (2020).

Question: What are the advantages of nonparametric smoothing methods?

They require few assumptions or restrictions on the data generating process. In particular, they do not assume a specific functional form for the function of interest (of course certain smoothness condition such as differentiability is required). They can deliver a consistent estimator for the unknown function, no matter whether it is linear or nonlinear. Thus, nonparametric methods can effectively reduce potential systematic biases due to model misspecification, which is more likely to be encountered for parametric modeling.

Question: What are the disadvantages of nonparametric methods?

- Nonparametric methods require a large data set for reasonable estimation. Furthermore, there exists a notorious problem of "curse of dimensionality," when the function of interest contains multiple explanatory variables. This will be explained below.

- There exists another notorious "boundary effect" problem for nonparametric estimation near the boundary regions of the support. This occurs due to asymmetric coverage of data in the boundary regions.

- Coefficients are usually difficult to interpret from an economic point of view.
- There exists a danger of potential overfitting, in the sense that nonparametric method, due to its flexibility, tends to capture non-essential features in a data which will not appear in out-of-sample scenarios.

The above two motivating examples are the so-called orthogonal series expansion methods. There are other nonparametric methods, such as splines smoothing, kernel smoothing, k-near neighbor, and local polynomial smoothing. As mentioned earlier, series expansion methods are examples of so-called **global smoothing**, because the coefficients are estimated using all observations, and they are then used to evaluate the values of the underlying function over all points in the support of X_t. A nonparametric series model is an increasing sequence of parametric models, as the sample size T grows. In this sense, it is also called a sieve estimator. In contrast, kernel and local polynomial methods are examples of the so-called **local smoothing** methods, because estimation only requires the observations in a neighborhood of the point of interest. Below we will mainly focus on kernel and local polynomial smoothing methods, due to their simplicity and intuitive nature.

2 Kernel Density Method

2.1 Univariate Density Estimation

Suppose $\{X_t\}$ is a strictly stationary time series process with unknown marginal PDF $g(x)$.

Question: How to estimate the marginal PDF $g(x)$ of the time series process $\{X_t\}$?

We first consider a parametric approach. Assume that $g(x)$ is an $N(\mu, \sigma^2)$ PDF with unknown μ and σ^2. Then we know the functional form of $g(x)$ up to two unknown parameters $\theta = (\mu, \sigma^2)'$:

$$g(x, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \qquad -\infty < x < \infty.$$

To estimate $g(x,\theta)$, it suffices to estimate two unknown parameters μ and σ^2. Based on the random sample $\{X_t\}_{t=1}^T$, we can obtain the maximum likelihood estimators (MLE),

$$\begin{aligned}\hat{\mu} &= \frac{1}{T}\sum_{t=1}^{T} X_t, \\ \hat{\sigma}^2 &= \frac{1}{T}\sum_{t=1}^{T}(X_t-\hat{\mu})^2.\end{aligned}$$

The approach taken here is called a parametric approach, that is, assuming that the unknown PDF is a known functional form up to some unknown parameters. It can be shown that the parameter estimator $\hat{\theta}$ converges to the unknown parameter value θ_0 at a root-T convergence rate in the sense that $\sqrt{T}(\hat{\theta}-\theta_0) = O_P(1)$, or $\hat{\theta}-\theta_0 = O_P(T^{-1/2})$, where $\hat{\theta} = (\hat{\mu},\hat{\sigma}^2)'$, $\theta_0 = (\mu_0,\sigma_0^2)'$, and $O_P(1)$ denotes boundedness in probability. The root-T convergence rate is called the parametric convergence rate for $\hat{\theta}$ and $g(x,\hat{\theta})$. As we will see below, nonparametric density estimators will have a slower convergence rate.

Question: What is the definition of $O_P(\delta_T)$?

Let $\{\delta_T, T \geq 1\}$ be a sequence of positive numbers. A random variable Y_T is said to be at most of order δ_T in probability, written $Y_T = O_P(\delta_T)$, if the sequence $\{Y_T/\delta_T, T \geq 1\}$ is tight, that is, if

$$\lim_{\lambda\to\infty}\limsup_{T\to\infty} P\left(|Y_T/\delta_T| > \lambda\right) = 0.$$

Tightness is usually indicated by writing $Y_T/\delta_T = O_P(1)$.

Question: What is the advantage of the parametric approach?

By the mean-value theorem, we obtain

$$\begin{aligned}g(x,\hat{\theta}) - g(x) &= g(x,\theta_0) - g(x) + \frac{\partial}{\partial\theta}g(x,\bar{\theta})(\hat{\theta}-\theta_0) \\ &= 0 + \frac{1}{\sqrt{T}}\frac{\partial}{\partial\theta}g(x,\bar{\theta})\sqrt{T}(\hat{\theta}-\theta_0) \\ &= 0 + O_P(T^{-1/2}) \\ &= O_P(T^{-1/2}).\end{aligned}$$

Intuitively, the first term, $g(x,\theta_0) - g(x)$, is the bias of the density estimator $g(x,\hat{\theta})$, which is zero if the assumption of correct model specification holds. The second term, $\frac{\partial}{\partial\theta}g(x,\bar{\theta})(\hat{\theta}-\theta_0)$, is due to the sampling error of the estimator $\hat{\theta}$, which is unavoidable no

matter whether the density estimator $g(x, \hat{\theta})$ is correctly specified. This term converges to zero in probability at the parametric root-T rate.

Question: What happens if the correct model specification assumption fails? That is, what happens if $g(x, \theta) \neq g(x)$ for all θ?

When the density model $g(x, \theta)$ is not correctly specified for the unknown PDF $g(x)$, the estimator $g(x, \hat{\theta})$ will not be consistent for $g(x)$ because the bias $g(x, \theta^*) - g(x)$ never vanishes no matter how large the sample size T is, where $\theta^* = p \lim \hat{\theta}$.

We now introduce a nonparametric estimation method for $g(x)$ which will not assume any restrictive functional form for $g(x)$. Instead, it lets data speak for the correct functional form for $g(x)$.

2.1.1 Kernel Density Estimator

Kernel smoothing is a kind of local smoothing. The purpose of nonparametric probability density estimation is to construct an estimate of a PDF without imposing restrictive functional form assumptions. Typically the only condition imposed on the unknown PDF is that it has at least first two order bounded derivatives. In this circumstance, we may use only local information about the value of the PDF at any given point in the support. That is, the value of the PDF of a point x must be calculated from data values that lie in a neighborhood of x, and to ensure consistency the neighborhood must shrink to zero as the sample size T increases. In the case of kernel density estimation, the radius of the effective neighborhood is roughly equal to the so-called "bandwidth" of a kernel density estimator, which is essentially a smoothing parameter. Under the assumption that the PDF is univariate with at least first two order bounded derivatives, and using a nonnegative kernel function, the size of bandwidth that optimizes the performance of the estimator in term of the mean squared error (MSE) criterion is proportional to the rate $T^{-1/5}$. The number of "parameters" needed to model the unknown PDF within a given interval is approximately equal to the number of bandwidths that can be fitted into that interval, and so is roughly of size $T^{1/5}$. Thus, nonparametric density estimation involves the adaptive fitting of approximately $T^{1/5}$ parameters, with this number growing with the sample size T.

Suppose we are interested in estimating the value of the PDF $g(x)$ at a given point x in the support of X_t. There are two basic instruments in kernel estimation: the kernel function $K(\cdot)$ and the bandwidth h. Intuitively, the former gives weighting to the observations in an interval containing the point x, and the latter controls the size of the interval containing observations.

We first introduce an important instrument for local smoothing. This is called a kernel function.

Definition [Second Order Kernel $K(\cdot)$]: A second order or positive kernel function $K(\cdot)$ is a pre-specified symmetric PDF such that

(1) $\int_{-\infty}^{\infty} K(u)du = 1$;
(2) $\int_{-\infty}^{\infty} K(u)udu = 0$;
(3) $\int_{-\infty}^{\infty} u^2K(u)du = C_K < \infty$;
(4) $\int_{-\infty}^{\infty} K^2(u)du = D_K < \infty$.

Intuitively, the kernel function $K(\cdot)$ is a weighting function that will "discount" the observations whose values are more away from the point x of interest.

The kernel functions satisfying the above condition are called a second order or positive kernel. It should be emphasized that the kernel $K(\cdot)$ has nothing to do with the unknown PDF $g(x)$ of $\{X_t\}$; it is just a weighting function for observations when constructing a kernel density estimator. More generally, we can define a q-th order kernel $K(\cdot)$, where $q \geq 2$.

Definition [qth Order Kernel]: $K(\cdot)$ satisfies the conditions that

(1) $\int_{-\infty}^{\infty} K(u)du = 1$;
(2) $\int_{-\infty}^{\infty} u^jK(u)du = 0$ for $1 \leq j \leq q-1$;
(3) $\int_{-\infty}^{\infty} u^qK(u)du < \infty$;
(4) $\int_{-\infty}^{\infty} K^2(u)du < \infty$.

For a higher order kernel (i.e., $q > 2$), $K(\cdot)$ will take some negative values at some points.

Question: Why is a higher order kernel useful? Can you give an example of a third order kernel? And an example of a fourth order kernel?

Higher order kernels can reduce the bias of a kernel estimator to a higher order. An example of higher order kernels is given in Robinson (1991).

We now consider some examples of second order kernels:

- Uniform kernel
$$K(u) = \frac{1}{2}\mathbf{1}(|u| \leq 1);$$
- Gaussian kernel
$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2); \qquad -\infty < u < \infty.$$
- Epanechnikov Kernel
$$K(u) = \frac{3}{4}(1-u^2)\mathbf{1}(|u| \leq 1);$$
- Quartic kernel
$$K(u) = \frac{15}{16}(1-u^2)^2\mathbf{1}(|u| \leq 1).$$

Among these kernels, the Gaussian kernel has unbounded support, while all other kernels have bounded supports of $[-1, 1]$. Also, the uniform kernel assigns an equal weighting within its support; in contrast, all other kernels have a downward weighting scheme.

Question: How does the kernel method work?

Let x be a fixed point in the support of X_t. Given a pre-chosen second kernel $K(u)$, we define a kernel density estimator for $g(x)$ based on the random sample $\{X_t\}_{t=1}^T$:

$$\begin{aligned}
\hat{g}(x) &= T^{-1}\sum_{t=1}^{T} K_h(x - X_t) \\
&= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{h} K\left(\frac{x - X_t}{h}\right) \\
&= \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right) d\hat{F}(y),
\end{aligned}$$

where
$$K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right),$$
$h = h(T) > 0$ is called a bandwidth or a window size, and $\hat{F}(y) = T^{-1}\sum_{t=1}^{T}\mathbf{1}(X_t \leq y)$ is the marginal empirical distribution function of the random sample $\{X_t\}_{t=1}^T$. This is exactly the same as the estimator introduced in Chapter 3, and it was first proposed by Rosenblatt (1956) and Parzen (1962) and so is also called the Rosenblatt-Parzen kernel density estimator.

We see immediately that the well-known histogram is a special case of the kernel density estimator $\hat{g}(x)$ with the choice of a uniform kernel.

Example 1 [Histogram]: If $K(u) = \frac{1}{2}\mathbf{1}(|u| \leq 1)$, then

$$\hat{g}(x) = \frac{1}{2hT}\sum_{t=1}^{T}\mathbf{1}(|x - X_t| \leq h).$$

Intuitively, with the choice of a uniform kernel, the kernel density estimator $\hat{g}(x)$ is the relative sample frequency of the observations on the interval $[x-h, x+h]$ which centers at point x and has a size of $2h$. Here, $2hT$ is approximately the sample size of the small interval $[x-h, x+h]$, when the size $2h$ is small enough. Alternatively, $T^{-1}\sum_{t=1}^{T}\mathbf{1}(|x - X_t| \leq h)$ is the relative sample frequency for the observations falling into the small interval $[x-h, x+h]$, which, by the law of large numbers, is approximately equal to the probability

$$\begin{aligned} E\left[\mathbf{1}(|x - X_t| \leq h)\right] &= P\left(x - h \leq X_t \leq x + h\right) \\ &= \int_{x-h}^{x+h} g(y)dy \\ &\approx 2hg(x) \end{aligned}$$

if h is small enough and $g(x)$ is continuous around the point x. Thus, the histogram is a reasonable estimator for $g(x)$, and indeed it is a consistent estimator $g(x)$ if h vanishes to zero but at a slower rate than sample size T goes to infinity.

Question: Under what conditions will the density estimator $\hat{g}(x)$ be consistent for the known density function $g(x)$?

We impose an assumption on the data generating process and the unknown PDF $g(x)$.

Assumption 3.1 [Smoothness of PDF]: (i) $\{X_t\}$ is a strictly stationary process with marginal *PDF* $g(x)$; (ii) $g(x)$ has a bounded support on $[a, b]$, and is continuously twice differentiable on $[a, b]$, with $g''(\cdot)$ being Lipschitz-continuous in the sense that $|g''(x_1) - g''(x_2)| \leq C|x_1 - x_2|$ for all $x_1, x_2 \in [a, b]$, where a, b and C are finite constants.

Question: How to define the derivatives at the boundary points?

By convention, the derivatives of $g(\cdot)$ at boundary points a and b are

$$\begin{aligned} g'(a) &= \lim_{x \to 0^+} \frac{g(a+x) - g(a)}{x}, \\ g'(b) &= \lim_{x \to 0^-} \frac{g(b+x) - g(b)}{x}. \end{aligned}$$

Similarly for the second derivatives $g''(a)$ and $g''(b)$ at the boundary points of the support $[a, b]$.

For convenience, we further impose an additional condition on kernel $K(\cdot)$, which will actually be maintained throughout this chapter.

Assumption 3.2 [Second Order Kernel with Bounded Support]: $K(u)$ is a positive kernel function with a bounded support on $[-1, 1]$.

This bounded support assumption is not necessary, but it simplifies the asymptotic analysis and interpretation.

2.1.2 Asymptotic Bias and Boundary Effect

Our purpose is to show that $\hat{g}(x)$ is a consistent estimator for $g(x)$ for a given point x in the support. Now we decompose

$$\hat{g}(x) - g(x) = [E\hat{g}(x) - g(x)] + [\hat{g}(x) - E\hat{g}(x)].$$

It follows that the mean squared error of the kernel density estimator $\hat{g}(x)$ is given by

$$\begin{aligned} \text{MSE}(\hat{g}(x)) &= [E\hat{g}(x) - g(x)]^2 + E\left[\hat{g}(x) - E\hat{g}(x)\right]^2 \\ &= \text{Bias}^2[\hat{g}(x)] + \text{var}\left[\hat{g}(x)\right]. \end{aligned}$$

The first term is the squared bias of the estimator $\hat{g}(x)$, which is nonstochastic, and the second term is the variance of $\hat{g}(x)$ at the point x. We shall show that under suitable regularity conditions, both the bias and the variance of $\hat{g}(x)$ vanish to zero as the sample size T goes to infinity.

We first consider the bias. For any given point x in the interior region $[a + h, b - h]$

of the support $[a, b]$ of X_t, we have

$$\begin{aligned}
E\left[\hat{g}(x)\right] - g(x) &= \frac{1}{T}\sum_{t=1}^{T} EK_h(x - X_t) - g(x) \\
&= E\left[K_h\left(x - X_t\right)\right] - g(x) \text{ (by identical distribution)} \\
&= \int_a^b \frac{1}{h} K\left(\frac{x-y}{h}\right) g(y)dy - g(x) \\
&= \int_{(a-x)/h}^{(b-x)/h} K(u)g(x+hu)du - g(x) \text{ (by change of variable } \frac{y-x}{h} = u) \\
&= \int_{-1}^{1} K(u)g(x+hu)du - g(x) \\
&= g(x)\int_{-1}^{1} K(u)du - g(x) \\
&\quad + hg'(x)\int_{-1}^{1} uK(u)du \\
&\quad + \frac{1}{2}h^2\int_{-1}^{1} u^2K(u)g''(x+\lambda hu)du \\
&= \frac{1}{2}h^2C_Kg''(x) + \frac{1}{2}h^2\int_{-1}^{1}[g''(x+\lambda hu) - g''(x)]u^2K(u)du \\
&= \frac{1}{2}h^2C_Kg''(x) + o(h^2)
\end{aligned}$$

where the second term

$$\int_{-1}^{1}[g''(x+\lambda hu) - g''(x)]u^2K(u)du \to 0$$

as $h \to 0$ by Lebesgue's dominated convergence theorem, and the boundedness and continuity of $g''(\cdot)$ and $\int_{-1}^{1} u^2K(u)du < \infty$.

Therefore, for the point x in the interior region $[a + h, b - h]$, the bias of $\hat{g}(x)$ is proportional to h^2. Thus, we must let $h \to 0$ as $T \to \infty$ in order to have the bias vanish to zero as $T \to \infty$.

The above result for the bias is obtained under the identical distribution assumption on $\{X_t\}$. It is irrelevant to whether $\{X_t\}$ is IID or serially dependent. In other words, it is robust to serial dependence in $\{X_t\}$.

Question: What happens to the bias of $\hat{g}(x)$ if x is outside the interior region $[a+h, b-h]$?

We say that x is outside the interior region $[a+h, b-h]$ if x in $[a, a+h]$ or $[b-h, b]$. These two regions are called boundary regions of the support. Their sizes are equal to h and so vanish to zero as the sample size T increases.

Suppose $x = a + \lambda h \in [a, a+h)$, where $\lambda \in [0, 1)$. We shall call x is a point in the left boundary region of the support $[a, b]$. Then

$$\begin{aligned}
E\left[\hat{g}(x)\right] - g(x) &= E\left[K_h\left(x - X_t\right)\right] - g(x) \\
&= \frac{1}{h}\int_a^b K\left(\frac{x-y}{h}\right) g(y)dy - g(x) \\
&= \int_{(a-x)/h}^{(b-x)/h} K(u)g(x+hu)du - g(x) \\
&= \int_{-\lambda}^{1} K(u)g(x+hu)du - g(x) \\
&= g(x)\int_{-\lambda}^{1} K(u)du - g(x) \\
&\quad + h\int_{-\lambda}^{1} uK(u)g'(x+\tau hu)du \\
&= g(x)\left[\int_{-\lambda}^{1} K(u)dx - 1\right] + O(h). \\
&= O(1)
\end{aligned}$$

if $g(x)$ is bounded away from zero, that is, if $g(x) \geq \epsilon > 0$ for all $x \in [a, b]$ for any small but fixed constant ϵ. Note that the $O(1)$ term arises since $\int_{-\lambda}^{1} K(u)dx = 1$ for any $\lambda < 1$.

Thus, if $x \in [a, a+h)$ or $(b-h, b]$, the bias $E[\hat{g}(x)] - g(x)$ may never vanish to zero even if $h \to 0$. This is due to the fact that there is no symmetric coverage of observations in the boundary region $[a, a+h)$ or $(b-h, b]$. This phenomenon is called the boundary effect or boundary problem of kernel estimation.

There have been several solutions proposed in the smoothed nonparametric literature. These include the following methods.

- **Trimming Observations:** Do not use the estimate $\hat{g}(x)$ when x is in the boundary regions. That is, only estimate and use the densities for points in the interior region $[a+h, b-h]$.

  This approach has a drawback. Namely, valuable information may be lost because $\hat{g}(x)$ in the boundary regions contain the information on the tail distribution of

$\{X_t\}$, which is particularly important to financial economists (e.g., extreme downside market risk) and welfare economics (e.g., the low-income population).

- **Using a Boundary Kernel:**

  To modify the kernel $K[(x - X_t)/h]$ when (and only when) x is the boundary regions such that it becomes location-dependent in the boundary region. For example, Hong and Li (2005) use a simple kernel-based density estimator

$$\hat{g}(x) = \frac{1}{T}\sum_{t=1}^{T} K_h(x, X_t),$$

  where

$$K_h(x,y) \equiv \begin{cases} h^{-1}K\left(\frac{x-y}{h}\right) / \int_{-(x/h)}^{1} K(u)du, & \text{if} \quad x \in [0,h), \\ h^{-1}K\left(\frac{x-y}{h}\right), & \text{if} \quad x \in [h, 1-h], \\ h^{-1}K\left(\frac{x-y}{h}\right) / \int_{-1}^{(1-x)/h} K(u)du, & \text{if} \quad x \in (1-h, 1] \end{cases}$$

  and $K(\cdot)$ is a standard second order kernel. The idea is to modify the kernel function in the boundary regions so that the integral of the kernel function is unity. Then the bias is $O(h^2)$ for all $x \in [a+h, b-h]$ in the interior region and is at most $O(h)$ for $x \in [a, a+h)$ and $(b-h, b]$ in the boundary regions. The advantage of this method is that it is very simple and always gives positive density estimates. The drawback is that the bias at the boundary region can be as slow as $O(h)$, which is slower than $O(h^2)$ in the interior region.

- **Using a Jackknife Kernel:** For x in the interior region $[a+h, b-h]$, use the standard positive kernel $K(\cdot)$. For x in the boundary regions $[a, a+h)$ and $(b-h, b]$, use the following jackknife kernel

$$K_\xi(u) \equiv (1+r)\frac{K(u)}{\omega_K(0,\xi)} - (r/\alpha)\frac{K(u/\alpha)}{\omega_K(0,\xi/\alpha)},$$

  where $\omega_K(l,\xi) \equiv \int_{-\xi}^{1} u^l K(u)du$ for $l = 0, 1$, $r \equiv r(\xi)$ and $\alpha \equiv \alpha(\xi)$ depend on parameter $\xi \in [0,1]$. When $x \in [a, a+h)$, we have $\xi = (x-a)/h$; when $x \in (b-h, b]$, we have $\xi = (b-x)/h$. In both cases, we set

$$r \equiv \frac{\omega_K(1,\xi)/\omega_K(0,\xi)}{\alpha\omega_K(1,\xi/\alpha)/\omega_K(0,\xi/\alpha) - \omega_K(1,b)/\omega_K(0,\xi)}.$$

As suggested in Rice (1986), we set $\alpha = 2 - \xi$. Given $\xi \in [0,1]$, the support of $K_\xi(\cdot)$ is $[-\alpha, \alpha]$. Consequently, for any $\xi \in [0,1]$,

$$\begin{aligned}
\int_{-\alpha}^{\alpha\xi} K_\xi(u)du &= \int_{-\alpha\xi}^{\alpha} K_\xi(u)du = 1, \\
\int_{-\alpha}^{\alpha\xi} uK_\xi(u)du &= -\int_{-\alpha\xi}^{\alpha} K_\xi(u)du = 0, \\
\int_{-\alpha}^{\alpha b} u^2 K_\xi(u)du &= \int_{-\alpha b}^{\alpha} u^2 K_\xi(u)du > 0, \\
\int_{-\alpha}^{\alpha b} K_\xi^2(u)du &= \int_{-\alpha b}^{\alpha} K_\xi^2(u)du > 0.
\end{aligned}$$

The bias is $O(h^2)$ for all points $x \in [a,b]$, including those in the boundary regions. We note that the jackknife kernel formula in Härdle (1990, Section 4.4) is incorrect.

- **Data Reflection:**

  The reflection method is to construct the kernel density estimate based on an augmented data which combined both the "reflected" data $\{-X_t\}_{t=1}^T$ and the original data $\{X_t\}_{t=1}^T$ with support on $[0,1]$. Suppose x is a boundary point in $[0,h)$ and $x \geq 0$. Then the reflection method gives an estimator

$$\hat{g}(x) = \frac{1}{T}\sum_{t=1}^{T} K_h(x - X_t) + \frac{1}{T}\sum_{t=1}^{T} K_h[x - (-X_t)].$$

  Note that with the support $[-1,1]$ of kernel $K(\cdot)$, when x is away from the boundary, the second term will be zero. Hence, this method only corrects the density estimate in the boundary region. See Schuster (1985, *Communications in Statistics: Theory and Methods*) and Hall and Wehrly (1991, *Journal of American Statistical Association*). This method has been extended by Chen and Hong (2012) and Hong, Sun and Wang (2018) to estimate time-varying functions (i.e., deterministic functions of time).

  Question: What is the general formula for the kernel density estimator when the support of X_t is $[a,b]$ rather than $[0,1]$?

  Suppose x is a boundary point in $[a, a+h)$.Then the density estimator becomes

$$\hat{g}(x) = \frac{1}{T}\sum_{t=1}^{T} K_h(x - X_t) + \frac{1}{T}\sum_{t=1}^{T} K_h[x - (-(X_t - a))].$$

- **Transformation:**

  Put $Y_t = q(X_t)$, where $q(\cdot)$ is a monotonic increasing function whose values range from $-\infty$ to ∞. Then

  $$\hat{g}_X(x) = q'(x)\hat{g}_Y[q(x)],$$

  where $\hat{g}_Y(\cdot)$ is the kernel density estimator for Y_t based on the transformed sample $\{Y_t\}_{t=1}^T$, which has infinite support.

  Question: Is there any free lunch with the use of the transformation method?

- **Local Polynomial Fitting.**

  Local polynomial automatically adapts to the boundary regions and the bias in the boundary region is the same in order of magnitude as the bias in the interior region. This will be discussed later.

2.1.3 Asymptotic Variance

Question: We have dealt with the bias of $\hat{g}(x)$. Then, what is the variance of $\hat{g}(x)$?

For the time being, in order to simplify the analysis, we assume an IID random sample. We will explain that the asymptotic result for the variance of $\hat{g}(x)$ remains true when $\{X_t\}$ is not IID under certain regularity restrictions on temporal dependence in $\{X_t\}$.

Assumption A.3 [IID Observations]: The random sample $\{X_t\}_{t=1}^T$ is IID.

The IID assumption simplifies our calculating the asymptotic variance of $\hat{g}(x)$. Later, we can relax the independence assumption for $\{X_t\}$ such that $\{X_t\}$ is an α-mixing process, a condition that allows weak temporal dependence (see Chapter 2). This will not change the asymptotic variance result for $\hat{g}(x)$.

Given any point x in the support $[a, b]$, put

$$Z_t \equiv Z_t(x) = K_h(x - X_t) - E\left[K_h(x - X_t)\right].$$

Then $\{Z_t\}_{t=1}^T$ is IID with mean zero. It follows that the variance of $\hat{g}(x)$,

$$\begin{aligned}
E\left[\hat{g}(x) - E\hat{g}(x)\right]^2 &= E\left(T^{-1}\sum_{t=1}^{T} Z_t\right)^2 \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\text{var}(Z_t) \\
&= \frac{1}{T}\text{var}(Z_t) \\
&= \frac{1}{T}\left[E[K_h^2(x-X_t)] - [EK_h(x-X_t)]^2\right] \\
&= \frac{1}{Th^2}\int_a^b K^2\left(\frac{x-y}{h}\right)g(y)dy \\
&\quad -\frac{1}{T}\left[\frac{1}{h}\int_a^b K\left(\frac{x-y}{h}\right)g(y)dy\right]^2 \\
&= \frac{1}{Th}g(x)\int_{-1}^{1} K^2(u)du[1+o(1)] + O(T^{-1}) \\
&= \frac{1}{Th}g(x)D_k + o(T^{-1}h^{-1}),
\end{aligned}$$

where the last second equality follows by change of variable $\frac{x-y}{h} = u$.

The variance of $\hat{g}(x)$ is proportional to $(Th)^{-1}$, which is the approximate sample size for the observations which fall into the interval $[x-h, x+h]$.

Next, we discuss the impact of serial dependence in $\{X_t\}$ on the asymptotic variance of $\hat{g}(x)$.

Question: What happens to the variance of $\hat{g}(x)$ if $\{X_t\}$ is serially dependent.

Suppose $\{X_t\}$ is a strictly stationary α-mixing process. Then under suitable conditions on the α-mixing coefficient $\alpha(j)$, for example, $\alpha(j) \leq Cj^{-\beta}$ for $\beta > \frac{5}{2}$, we have the same MSE formula for $\hat{g}(x)$ as we have when $\{X_t\}$ is IID. This is formally established in Robinson (1983). See Robinson (1983, *Journal of Time Series Analysis*) for details. Intuitively, when a bounded support kernel $K(u)$ is used, the kernel density estimator is an weighted average of nonlinear functions of observations $\{X_t\}_{t=1}^T$ which fall into the small interval $[x-h, x+h]$. The observations that fall into the small interval are determined by the closeness of their values to the value of x, not by closeness in time. Suppose we re-lable the observations falling into the small interval by a subsequence or subsample $\{\tilde{X}_{t^*}\}_{t^*=1}^{T^*}$. Then the size of the subsample $T^*$is of the order of Th, and the

time index t^* are not consercative time periods but rather far away from each other in time. This implies that the observations in the subsequence behave like an IID sequence when serial dependence in the original time series $\{X_t\}$ is not too strong.

We now formally examine the impact of serial dependence of $\{X_t\}$ on the asymptotic variance of the kernel density estimator $\hat{g}(x)$. For this aim, we first introduce the strong mixing condition which is called α-mixing.

Definition [α-mixing]: Let $\{X_t\}$ be a strictly stationary time series process. For $j = 1, 2, ...,$ define

$$\alpha(j) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_j^\infty} |P(A \cap B) - P(A)P(B)|,$$

where $\mathcal{F}_i^j$ denotes the σ-algebra generated by $\{X_t, i \leq t \leq j\}$. Then the process $\{X_t\}$ is said to be α-mixing if $\alpha(j) \to 0$ as $j \to \infty$.

Similar to ergodicity, the α-mixing condition is a concept for asymptotic independence. A mixing process can be viewed as a sequence of random variables for which the past and distant future are asymptotically independent. The α-mixing condition implies ergodicity. See White (*Asymptotic Theory for Econometricians*, 2001). In fact, there are several concepts of mixing, such as α-mixing, β-mixing, and ϕ-mixing. Among them, α-mixing is the weakest condition on serial dependence; it is also called strong mixing.

If $\{X_t\}$ is a strictly stationary Markov chain, the mixing coefficient $\alpha(j)$ can be effectively defined with $(\mathcal{F}_{-\infty}^0, \mathcal{F}_j^\infty)$ replaced by $(\sigma(X_0), \sigma(X_n))$, and in this case,

$$\alpha(j) \leq \frac{1}{2} \int \int |f_j(x, y) - g(x)g(y)| \, dxdy,$$

where $f_j(x, y)$ is the joint PDF of (X_t, X_{t-j}).

We first state a useful lemma.

Lemma [Doukhan (1994)]: Let X and Y be two real random variables. Define

$$\alpha = \sup_{A \in \sigma(X), B \in \sigma(Y)} |P(A \cap B) - P(A)P(B)|.$$

(1) Suppose $E(|X|^p + |X|^q) < \infty$ for some $p, q \geq 1$, and $1/p + 1/q < 1$. Then

$$|\text{cov}(X, Y)| \leq 8\alpha^{1/r} (E|X|^p)^{1/p} (E|X|^q)^{1/q},$$

where $r = (1 - 1/p - 1/q)^{-1}$.
(2) If $P(|X| \leq C_1) = 1$ and $P(|Y| \leq C_2) = 1$ for some constants C_1 and C_2, then

$$|\text{cov}(X, Y)| \leq 4\alpha C_1 C_2.$$

Theorem [Asymptotic Variance of $\hat{g}(x)$ under Mixing Conditions]: Let $\{X_t\}$ be a strictly stationary α-mixing process with the mixing coefficient $\alpha(j) \leq C_j^{-\beta}$ for some $C > 0$ and $\beta > 2$. Assume that the pairwise joint density function $f_j(x, y)$ is bounded uniformly in (x, y) and in lag order j. Then for $x \in [a, b]$, where a, b are finite constants,

$$\text{var}\left[\hat{g}(x)\right] = \frac{1}{Th} g(x) D_K + o(T^{-1}h^{-1}).$$

Proof: Put

$$Z_t = K_h(x, X_t) = \frac{1}{h} K(\frac{x - X_t}{h}).$$

Then by strict stationarity of $\{X_t\}$, we have

$$\begin{aligned} \text{var}\left[\hat{g}(x)\right] &= \text{var}\left(T^{-1}\sum_{t=1}^{T} Z_t\right) \\ &= \frac{1}{T}\text{var}(Z_1) + 2\frac{1}{T}\sum_{j=1}^{T-1}(1 - j/T)\text{cov}(Z_0, Z_j). \end{aligned}$$

Note that $E(Z_1) = E[\hat{g}(x)] = O(1)$. By change of variable, we have

$$\begin{aligned} \text{var}(Z_1) &= E[K_h^2(x, X_1)] - (EZ_1)^2 \\ &= \frac{1}{h} g(x) D_K + O(1). \end{aligned}$$

where the first term dominates the second term in terms of order of magnitude. Thus, it remains to show

$$\frac{1}{T}\sum_{j=1}^{T-1} \text{cov}(Z_0, Z_j) = o(h^{-1}).$$

Because $|Z_0| \leq Ch^{-1}$, we have

$$|\text{cov}(Z_0, Z_j)| \leq 4(Ch^{-1})^2 \alpha(j)$$

by Billingsley's inequality. It follows that

$$\sum_{j=m(T)+1}^{T-1} |\text{cov}(Z_0, Z_j)| \leq 4C^2 h^{-2} \sum_{j=m(T)+1}^{T-1} j^{-\beta} \leq C^3 m(T)^{1-\beta} h^{-2}$$

where $m(T) \to \infty$ as $T \to \infty$.

On the other hand,

$$\begin{aligned} |\text{cov}(Z_0, Z_j)| &= |E(Z_0 Z_j) - E(Z_0)E(Z_j)| \\ &\leq \int K_h(x, x')K_h(y, y')f_j(x', y')dx'dy' + [E(Z_0)]^2 \\ &\leq C\left[\int_{-\infty}^{\infty} K_h(x, x')dx'\right]^2 + [E(Z_0)]^2 \\ &\leq C^2. \end{aligned}$$

Hence, we have

$$\sum_{j=1}^{m(T)} |\text{cov}(Z_0, Z_j)| \leq Cm(T).$$

By setting $m(T) = h^{-2/\beta}$, we have

$$\sum_{j=1}^{T-1} |\text{cov}(Z_0, Z_j)| = O(h^{-2/\beta}) = o(h^{-1})$$

for $\beta > 2$. This completes the proof.

The asymptotic variance of $\hat{g}(x)$ is exactly the same as that under the IID assumption on $\{X_t\}$. This is a bit surprising, but it is true. It follows because when $\{X_t\}$ is not IID, the variance of $\hat{g}(x)$ can be decomposed as the sum of individual variances $\{\text{var}[K_h(x, X_t)]\}_{t=1}^T$ and the sum of all possible covariance terms $\{\text{cov}[K_h(x, X_t), K_h(x, X_s)]\}_{t \neq s}$ together. Given the strong-mixing condition, the sum of all possible covariance terms $\{\text{cov}[K_h(x, X_t), K_h(x, X_s)]\}_{t \neq s}$ together is of smaller order in magnitude than the sum of all individual variances $\{\text{var}[K_h(x, X_t)]\}_{t=1}^T$, due to the smoothing parameter h.

Hart (1996) provides a nice intuition for this result. Suppose the kernel $K(\cdot)$ has support on $[-1, 1]$, as assumed in this chapter. Then the kernel density estimator at the point x uses only the local data points inside the local interval $[x - h, x + h]$. The observations whose values fall into this local interval are generally far away from each other in *time*. Thus, although the data $\{X_t\}_{t=1}^T$ in the original sequence may be highly correlated, the dependence for the new subsequence in the local interval around x can be much weaker. As a result, the local data look like those from an independent sample. Hence, one would expect that the asymptotic variance of the kernel density estimator is the same as that for the independent observations when certain mixing conditions are imposed.

References: *Kernel estimation in time series*: Robinson (1983, *Journal of Time Series Analysis*), Fan and Yao (2003, *Nonlinear Time series*)

2.1.4 MSE and Optimal Bandwidth

It follows that the mean squared error (MSE) of $\hat{g}(x)$ is given by

$$\begin{aligned} MSE[\hat{g}(x)] &= E\left[\hat{g}(x) - g(x)\right]^2 \\ &= \text{var}[\hat{g}(x)] + \text{Bias}^2[\hat{g}(x), g(x)] \\ &= \frac{1}{Th} g(x) D_K + \frac{1}{4} h^4 \left[g''(x)\right]^2 C_K^2 + o(T^{-1}h^{-1} + h^4) \\ &= O(T^{-1}h^{-1} + h^4). \end{aligned}$$

By Chebyshev's inequality, for any given point x in the interior region $[a+h, b-h]$, we have

$$\hat{g}(x) - g(x) = O_P(T^{-1/2}h^{-1/2} + h^2).$$

Therefore, for $\hat{g}(x) \to^p g(x)$, we need $Th \to \infty, h \to 0$ as $T \to \infty$. Under the stated assumptions, the estimator $\hat{g}(x)$ is always consistent for the unknown density $g(x)$ but at a slower rate than the parametric $T^{-1/2}$. This means that a large sample is needed to obtain a reasonable estimate for $g(x)$.

Moreover, the bias of $\hat{g}(x)$ depends on the smoothness of the unknown function $g(\cdot)$. In particular, if the second derivative $g''(x)$ has a relatively sharp spike at the point x, it is difficult to obtain a good estimate $g(\cdot)$ at the point x.

We can also obtain a relative MSE criterion when $g(x) > 0$:

$$\begin{aligned} MSE[\hat{g}(x)/g(x)] &= \frac{MSE[\hat{g}(x)]}{g^2(x)} \\ &= E\left[\frac{\hat{g}(x) - g(x)}{g(x)}\right]^2 \\ &= \frac{1}{Thg(x)} D_K + \frac{1}{4} h^4 \left[\frac{g''(x)}{g(x)}\right]^2 C_K^2. \\ &\quad + o(T^{-1}h^{-1} + h^4) \\ &= O(T^{-1}h^{-1} + h^4) \end{aligned}$$

The expression of the relative MSE indicates that it is very difficult to obtain a reasonable estimate of $g(x)$ in the sparse area where relatively few observations are

available (i.e., when $g(x)$ is small), or in the area where $g(\cdot)$ changes dramatically (i.e., when the curvature $g''(x)/g(x)$ is large in absolute value).

As can be seen from the MSE formula for $\hat{g}(x)$, a small bandwidth h will reduce the bias but inflate the variance, and a large bandwidth will increase the bias but reduce the variance. The bandwidth is a smoothing parameter. When the bandwidth h is so small such that the squared bias is smaller than the variance, we say that there exists undersmoothing; when the bandwidth is so large such that its squared bias is larger than the variance, we say that there exists oversmoothing. Optimal smoothing is achieved if the bandwidth balances the squared bias and the variance of $\hat{g}(x)$. We now consider the optimal choice of the bandwidth h. The optimal bandwidth can be obtained by minimizing $MSE[\hat{g}(x)]$:

$$h_0 = \left[\frac{D_K}{C_K^2}\frac{1/g(x)}{[g''(x)/g(x)]^2}\right]^{\frac{1}{5}} T^{-1/5}.$$

The less smooth the PDF $g(x)$ is or the more sparse the observations are around the point x, the smaller the optimal bandwidth h_0 for any given sample size T. The optimal bandwidth h_0 gives the optimal convergence rate for $\hat{g}(x)$:

$$\hat{g}(x) - g(x) = O_P(T^{-2/5}).$$

The convergence rate $T^{-2/5}$ is slower than the parametric rate $T^{-1/2}$.

The optimal bandwidth h_0 is unknown, because it depends on the unknown density function $g(x)$ and its second order derivative $g''(x)$.

Question: How to obtain a consistent estimator of this optimal bandwidth in practice?

Since we have obtained a closed form expression for the optimal bandwidth h_0, we can obtain a consistent estimator of h_0 by plugging in some preliminary consistent estimators for $g(x)$ and $g''(x)$. Suppose we have some initial preliminary estimators, say $\tilde{g}(x)$ and $\tilde{g}''(x)$, for $g(x)$ and $g''(x)$ respectively. Then we can plug them into the above formula for h_0, obtaining an estimator for h_0. With such a data-dependent bandwidth, we obtain a new kernel estimator which has better statistical properties than an arbitrary choice of h. This is the well-known plug-in method. We note that even if $\tilde{g}(x)$ and $\tilde{g}''(x)$ are not consistent for $g(x)$ and $g''(x)$, then the second stage kernel density estimator $\hat{g}(x)$ is still consistent for $g(x)$, although it is not optimal. However, consistency of $\tilde{g}(x)$ and $\tilde{g}''(x)$

for $g(x)$ and $g''(x)$ respectively ensures that $\hat{g}(x)$ is an asymptotically optimal estimator for $g(x)$.

In addition to the plug-in method to choose a data-driven bandwidth, there exists other data-driven methods to choose an optimal bandwidth. One example is the cross-validation method. For more discussion, see xxx.

As can be seen from the MSE formula, both the variance and squared bias of $\hat{g}(x)$ depend on the kernel function K. We now consider the choice of an optimal kernel. Using the calculus of variation, it can be shown, as in Epanechnikov (1969, *Theory of Probability and Its Applications*), that the optimal kernel that minimizes the MSE of $\hat{g}(x)$ over a class of positive kernel functions is the so-called Epanechnikov kernel:

$$K(u) = \frac{3}{4}(1-u^2)\mathbf{1}(|u|<1).$$

In practice, it is found that the choice of h is more important than the choice of $K(u)$. See also Priestley (1962).

2.2 Multivariate Density Estimation

We now extend the kernel method to estimate a multivariate density function when X_t is a strictly stationary vector-valued time series process.

Question: How to estimate a joint PDF $f(x)$ of $X_t = (X_{1t}, X_{2t}, ..., X_{dt})'$, where $x = (x_1, x_2, ..., x_d)'$ is a $d \times 1$ vector?

Example 1: How to estimate the joint PDF $f_j(x, y)$ of (X_t, X_{t-j})?

To estimate $f(x)$, we define a product kernel density estimator

$$\begin{aligned}\hat{f}(x) &= \frac{1}{T}\sum_{t=1}^{T}\prod_{i=1}^{d} K_h(x_i - X_{it}) \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathcal{K}_h(x - X_t),\end{aligned}$$

where

$$\mathcal{K}_h(x - X_t) = \prod_{i=1}^{d} K_h(x_i - X_{it}).$$

For simplicity, we have used the same bandwidth h for every coordinate. Different bandwidths could be used for different coordinates. In practice, before using the same bandwidth h, one can stabdardize all $\{X_{it}\}_{t=1}^{T}$ by dividing by their sample standard deviations respectively for all $i = 1, .., d$.

We first consider the bias of $\hat{f}(x)$. Suppose x is an interior point x such that $x_i \in [a_i + h, b_i - h]$ for all $i = 1, ..., d$. This implies that x is in a d-dimensional box each side of which is $[a_i + h, b_i - h]$ for $i = 1, ..., d$. It follows that the bias of $\hat{f}(x)$,

$$\begin{aligned}
E\left[\hat{f}(x)\right] - f(x) &= E\mathcal{K}_h(x - X_t) - f(x) \\
&= E\prod_{i=1}^{d} K_h\left(x_i - X_{it}\right) - f(x) \\
&= \int \cdots \int \left[\prod_{i=1}^{d} \frac{1}{h} K\left(\frac{x_i - y_i}{h}\right)\right] f(y)dy - f(x) \\
&= \prod_{i=1}^{d} \int_{(a_i - x_i)/h}^{(b_i - x_i)/h} K(u_i) f(x + hu)du - f(x) \\
&= \int_{-1}^{1} \cdots \int_{-1}^{1} \prod_{i=1}^{d} K(u_i) f(x + hu)du - f(x) \\
&= f(x) \prod_{i=1}^{d} \int_{-1}^{1} K(u_i)du_i - f(x) \\
&\quad + h \sum_{i=1}^{d} f_i(x) \int_{-1}^{1} u_i K(u_i)du_i \\
&\quad + \frac{1}{2} h^2 \sum_{i=1}^{d} \sum_{j=1}^{d} \int_{-1}^{1} \int_{-1}^{1} u_i u_j K(u_i) K(u_j) f_{ij}(x + \lambda uh)du_i du_j \\
&= \frac{1}{2} h^2 C_K \sum_{i=1}^{d} f_{ii}(x) + o(h^2) \\
&= O(h^2).
\end{aligned}$$

where $f_i(x) = \frac{\partial}{\partial x_i} f(x)$, $f_{ij}(x) = \frac{\partial^2}{\partial x_i \partial x_j} f(x)$, and the quantity $\sum_{i=1}^{d} f_{ii}(x)$ is called the Laplace of the joint PDF $f(x)$.

Next, put

$$\begin{aligned} Z_t &\equiv Z_t(x) \\ &= \mathcal{K}_h(x - X_t) - E\mathcal{K}_h(x - X_t) \\ &= \prod_{i=1}^{d} K_h(x_i - X_{it}) - E\prod_{i=1}^{d} K_h(x - X_{it}). \end{aligned}$$

Then $\{Z_t\}$ is IID with mean zero given that $\{X_t\}$ is IID. It follows that the variance of $\hat{f}(x)$

$$\begin{aligned} E\left[\hat{f}(x) - E\hat{f}(x)\right]^2 &= E\left[T^{-1}\sum_{t=1}^{T}[\mathcal{K}_h(x - X_t) - E\mathcal{K}_h(x - X_t)]\right]^2 \\ &= \frac{1}{T^2}\sum_{t=1}^{T} E(Z_t^2) \text{ (by independence)} \\ &= \frac{1}{T}E\left[\prod_{i=1}^{d} K_h(x_i - X_{it}) - E\prod_{i=1}^{d} K_h(x_i - X_{it})\right]^2 \\ &= \frac{1}{T}\left[E\prod_{i=1}^{d} K_h^2(x_i - X_{it}) - \left[E\prod_{i=1}^{d} K_h(x_i - X_{it})\right]^2\right] \\ &= \frac{1}{Th^d} f(x) D_K^d + o(T^{-1}h^{-d}). \end{aligned}$$

We note that the asymptotic variance of $\hat{f}(x)$ is proprotional to the inverse of Th^d, where Th^d is approximately the effective sample size for observations falling into a d-dimensional subspace centered at point x, with each size equal to $2h$. The asymptotic variance of $\hat{f}(x)$ remains valid under a suitale α-mixing condition for the time series $\{X_t\}$.

It follows that the MSE of $\hat{f}(x)$ is

$$\begin{aligned} & MSE[\hat{f}(x)] \\ = & \frac{1}{Th^d} f(x) D_K^d + \frac{1}{4} C_K^2 h^4 \left[\sum_{i=1}^{d} f_{ii}(x)\right]^2 \\ & + o(T^{-1}h^{-d} + h^4) \\ = & O(T^{-1}h^{-d} + h^4). \end{aligned}$$

With a suitable choice of bandwidth h, the optimal MSE convergence rate of $\hat{f}(x)$ to $f(x)$ is $T^{-\frac{4}{4+d}}$, which can be obtained by setting the bandwidth

$$h_0 = \left[\frac{dD_K^2}{C_K^2} \frac{1/f(x)}{[\sum_{i=1}^d f_{ii}(x)/f(x)]^2} \right]^{\frac{1}{d+4}} T^{-\frac{1}{d+4}}.$$

Thus, the MSE convergence rate is

- $\text{MSE}[\hat{f}(x)] \propto T^{-\frac{4}{5}}$ if $d = 1$,
- $\text{MSE}[\hat{f}(x)] \propto T^{-\frac{2}{3}}$ if $d = 2$,
- $\text{MSE}[\hat{f}(x)] \propto T^{-\frac{4}{7}}$ if $d = 3$.

The larger dimension d, the slower convergence of $\hat{f}(x)$. This is the so-called "**curse of dimensionality**" associated with multivariate nonparametric estimation. It implies that a large sample size is needed in order to have a reasonable estimation for $f(x)$. In particular, the same size T has to be increased exponentially fast as the dimension d increases in order to achieve the same level of estimation accuracy and precision. For most typical sample sizes encounterd in economics and finance, it is rare to see nonaparemetric estimation with dimension $d > 5$.

Question: How to deal with the curse of dimensionality?

There are various methods to deal with the curse of dimensionality. For example,

- **Imposing Multiplicability or Additicability Conditions.**

  Suppose multiplicability conditions for $f(x)$ holds such as

  $$f(x) = \prod_{i=1}^{d} g_i(x_i).$$

  Then one can estimate $g_i(x_i)$ separately. When $f(x)$ is a joint density function, multiplicability occurs when and only when $X_{1t}, X_{2t}, ..., X_{dt}$ are mutually indpendent. For functions other than probability densities, an additivity condition can be imposed to reduce the dimension of estimation.

- **Projection Pursuit.**

  This approach assumes that a multivariate function is an unknown function of the linear combination of d explanary variables, and then use a nonparametric method to estimate the unknown function and the combination coefficients. A well-known class of models in econometrics is single-index models for which a function of X_t is assumed to be an unknown function of a linear combination of the components of X_t, where the combination coefficients are also unknown.

- **Imposing the Markov Condition.**

  Suppose the time series $\{X_t\}$ is a Markov process. Then

$$\begin{aligned} f(X_t|I_{t-1}) &= f(X_t|X_{t-1}) \\ &= \frac{f(X_t, X_{t-1})}{g(X_{t-1})}, \end{aligned}$$

  where $I_{t-1} = (X_{t-1}, X_{t-2}, ...)$ is the set of infinite dimension. Here, $f(X_t, X_{t-1})$ depends on only X_t and X_{t-1}.

Kernel density estimators have been widely used in time series econometrics and financial econometrics. For example,

- Aït-Sahalia (1996, *Review of Financial Studies*) uses the kernel-based marginal density estimator $\hat{g}(x)$ to test the adequacy of a diffusion model for short-term interest rates.

- Gallant and Tauchen (1996, *Econometric Theory*) use the Hermite polynomial-based estimator for the conditional pdf of X_t given I_{t-1} to estimate continuous-time models efficiently.

- Hong and Li (2005, *Review of Financial Studies*) use the kernel-based joint density estimator $\hat{f}_j(x, y)$ to test the adequacy of continuous-time models and consider an application to affine term structure models of interest rates.

- Hong and White (2005, *Econometrica*) use the kernel-based joint density estimator $\hat{f}_j(x, y)$ to construct a nonparametric entropy-density measure for serial dependence with a well-defined asymptotic distribution.

- Su and White (2008, *Econometric Theory*) propose a Hellinger metric-based test for conditional dependence test which is applicable to test for general Granger causality by checking whether

$$f(X_t|X_{t-1},...,X_{t-p}) = f(X_t|X_{t-1},...,X_{t-p},Y_{t-1},...,Y_{t-q}),$$

  where the conditional PDFs are estimated using the kernel method, and the lag orders p and q are given in advance.

- de Matos and Fernandes (2007, *Journal of Econometrics*) propose a test for the Markov property of a time series process:

$$f(X_t|I_{t-1}) = f(X_t|X_{t-1}).$$

  They compare two kernel estimators for the conditional PDFs

$$f(X_t|X_{t-1},X_{t-j}) = \frac{f(X_t,X_{t-1},X_{t-j})}{f(X_{t-1},X_{t-j})}$$

  and

$$f(X_t|X_{t-1}) = \frac{f(X_t,X_{t-1})}{f(X_{t-1})}.$$

- Wang and Hong (2017, *Econometric Theory*) propose a test for conditional independence which is applicable to test the Markob property and Granger causality in distribution. They estimate the conditional characteristic functuon rather than the conditional density function of $\{X_t\}$, thus avoiding the curse of dimensionality problem when the dimension d of X_t is large.

There are other possible approaches to testing the Markov property.
In general, this requires checking whether

$$f(X_t = x|I_{t-1}) = f(X_t = x|X_{t-1}),$$

where $I_{t-1} = \{X_t, X_{t-1}, ...,\}$. A possible approach to testing the Markov property of a time series can be based on the following lemma.

Lemma [Probability Integral Transforms of Markov Process]: Suppose $\{X_t\}$ is a strictly stationary process. Denote the conditional PDF of X_t given X_{t-1} as

$$f(x|y) = f(X_t = x|X_{t-1} = y).$$

Define the probability integral transform

$$Z_t = \int_{-\infty}^{X_t} f(x|X_{t-1})dx = F_t(X_t),$$

where $F_t(x) = P(X_t \leq x|X_{t-1})$. If $\{X_t\}$ is Markovian, then

$$\{Z_t\} \sim \text{IID } U[0,1].$$

3 Nonparametric Regression Estimation

Question: How to estimate a regression function $E(Y_t|X_t)$ using an observed bivariate sample $\{Y_t, X_t\}_{t=1}^T$? Note that both Y_t and X_t are random variables here.

We first consider a few examples of regression functions.

Example 1: The autoregression function

$$r_j(X_{t-j}) = E(X_t|X_{t-j}).$$

We can write

$$X_t = r_j(X_{t-j}) + \varepsilon_t,$$

where $E(\varepsilon_t|X_{t-j}) = 0$.

Example 2: The conditional variance

$$\begin{aligned} \sigma_j^2(x) &= \text{var}(X_t|X_{t-j}) \\ &= E(X_t^2|X_{t-j}) - [E(X_t|X_{t-j})]^2 . \end{aligned}$$

Example 3: The conditional distribution function

$$\begin{aligned} F_t(x) &= P(X_t \leq x|I_{t-1}) \\ &= E\left[\mathbf{1}(X_t \leq x)|I_{t-1}\right], \end{aligned}$$

where I_{t-1} is an information set available at time $t-1$. If we assume that $\{X_t\}$ is an Markovian process. Then

$$F_t(x) = E\left[\mathbf{1}(X_t \leq x)|X_{t-1}\right].$$

This is a generalized regression function of $\mathbf{1}(X_t \leq x)$ on X_{t-1}.

Example 4: The conditional characteristic function

$$\varphi_t(u) = E\left[\exp(iuX_t)|I_{t-1}\right].$$

If $\{X_t\}$ is an Markovian process, then

$$\varphi_t(u) = E\left[\exp(iuX_t)|X_{t-1}\right].$$

This is a generalized regression function of $\exp(iuX_t)$ on X_{t-1}.

3.1 Kernel Regression Estimation

We first impose a regularity condition on the data generating process and the regression function.

Assumption [DGP]: (i) Suppose $\{Y_t, X_t\}'$ is an IID sequence such that the regression function $r(x) \equiv E(Y_t|X_t = x)$ exists and is twice continuously differentiable; (ii) X_t is a continuous random variable with support $[a, b]$ and probability density $g(x)$ which is also twice continuously differentiable over $[a, b]$. Furthermore, $g(x) > 0$ for all $x \in [a, b]$.

We will relax the IID assumption to a serially dependent time series process at a later stage. Like in the case of density estimation, allowing mild serial dependence (e.g., α-mixing) in $\{Y_t, X_t\}'$ will not affect the asymptotic results derived under the IID assumption.

Question: How to estimate the regression function $r(x)$?

We can always write

$$Y_t = r(X_t) + \varepsilon_t,$$

where $E(\varepsilon_t|X_t) = 0$ and $\text{var}(\varepsilon_t|X_t) = \sigma^2(X_t)$. Note that conditional heteroskedasticity may exist. We impose a continuity condition on $\sigma^2(x)$.

Assumption [Conditional Heteroskedasticity]: $\sigma^2(x)$ is continuous over the support $[a, b]$ of X_t.

3.1.1 Nadaraya-Watson Estimator

For any given x in the support of X_t, define a kernel-based regression estimator

$$\hat{r}(x) = \frac{\hat{m}(x)}{\hat{g}(x)},$$

where the numerator

$$\hat{m}(x) = \frac{1}{T}\sum_{t=1}^{T} Y_t K_h(x - X_t)$$

is a weighted sample mean of $\{Y_t\}$, and, as before, the denominator

$$\hat{g}(x) = T^{-1}\sum_{t=1}^{T} K_h(x - X_t)$$

is the kernel estimator for density $g(x)$ at point x. This kernel regression estimator was proposed by Nadaraya (1964) and Watson (1964) and so is also called the Nadaraya-Watson estimator.

Alternatively, we can express

$$\hat{r}(x) = \sum_{t=1}^{T} \hat{W}_t(x) Y_t,$$

where the weighting function

$$\hat{W}_t(x) = \frac{K_h(x - X_t)}{\sum_{t=1}^{T} K_h(x - X_t)},$$

which sums to unity, that is,

$$\sum_{t=1}^{T} \hat{W}_t = 1.$$

Therefore, the Nadaraya-Watson estimator is a local weighted sample mean of $\{Y_t\}_{t=1}^{n}$, where the weight $\hat{W}_t(x)$ is zero outside the interval $[x - h, x + h]$ when the kernel $K(u)$ has bounded support on $[-1, 1]$.

We first provide a geometric interpretation for $\hat{r}(x)$. When the uniform kernel $K(u) = \frac{1}{2}\mathbf{1}(|u| \leq 1)$ is used, the Nadaraya-Watson estimator becomes

$$\hat{r}(x) = \frac{\sum_{t=1}^{T} Y_t \mathbf{1}(|X_t - x| \leq h)}{\sum_{t=1}^{T} \mathbf{1}(|X_t - x| \leq h)}.$$

This is a local sample mean, that is, the average of the observations $\{Y_t\}_{t=1}^T$ for which the values of the corresponding explanatory variables $\{X_t\}$ fall into the interval $[x-h, x+h]$. Tukey (1961) calls this estimator the **regressogram**. Intuitively, suppose $r(\cdot)$ is a smooth function, and we consider a small interval $[x-h, x+h]$, which is centered at point x and has size $2h$, where h is small. Then $r(\cdot)$ will be nearly a constant over this small interval and can be estimated by taking an average of the observations $\{Y_t\}$ which correspond to those $\{X_t\}$ whose values fall into the small interval.

More generally, we can assign different weights to observations of $\{Y_t\}_{t=1}^T$ according to their distances to the location x. This makes sense because the observations $\{X_t\}_{t=1}^T$ closer to x will contain more information about $r(x)$ at point x. The use of $K(\cdot)$ is to assign different weights for observations $\{Y_t, X_t\}_{t=1}^T$.

Kernel regression is a special convolution filter used in engineering.

3.1.2 MSE and Optimal Bandwidth

Question: How to derive the asymptotic MSE of $\hat{r}(x)$?

The Nadaraya-Watson estimator $\hat{r}(x)$ is a ratio of two random variables. To simplify asymptotic analysis, for any given point x, we consider the decomposition

$$\begin{aligned}
\hat{r}(x) - r(x) &= \frac{\hat{m}(x) - r(x)\hat{g}(x)}{\hat{g}(x)} \\
&= \frac{\hat{m}(x) - r(x)\hat{g}(x)}{E[\hat{g}(x)]} \\
&\quad + \frac{[\hat{m}(x) - r(x)\hat{g}(x)]}{E[\hat{g}(x)]} \cdot \frac{[E[\hat{g}(x)] - \hat{g}(x)]}{\hat{g}(x)} \\
&= \frac{\hat{m}(x) - r(x)\hat{g}(x)}{E[\hat{g}(x)]} \\
&\quad + \text{higher order term.}
\end{aligned}$$

Here the second term is of a higher order term because

$$\hat{g}(x) - E[\hat{g}(x)] \xrightarrow{p} 0 \text{ as } T \to \infty$$

and

$$E[\hat{g}(x)] \to g(x) \int_{-1}^{1} K(u)du > 0 \text{ as } h \to 0.$$

It can be shown that the second term is of a higher order term that vanishes faster than the first term (**how?**). As a consequence, the convergence rate of $\hat{r}(x)$ to $r(x)$ is determined by the first term, which is the dominant term.

For the first term, using $Y_t = r(X_t) + \varepsilon_t$, we can write the numerator as follows:

$$\begin{aligned}
\hat{m}(x) - r(x)\hat{g}(x) &= \frac{1}{T}\sum_{t=1}^{T}[Y_t - r(x)]K_h(x - X_t) \\
&= \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t K_h(x - X_t) \\
&\quad + \frac{1}{T}\sum_{t=1}^{T}[r(X_t) - r(x)]K_h(x - X_t) \\
&= \hat{V}(x) + \hat{B}(x), \quad say, \\
&= \text{variance component + bias component.}
\end{aligned}$$

Here, the first term $\hat{V}(x)$ is a variance effect, and the second term $\hat{B}(x)$ is a bias effect.

We first consider the variance term. For simplicity, we first assume that $\{Y_t, X_t\}$ is an IID sequence. Then for the variance component, we have

$$\begin{aligned}
E\left[\hat{V}(x)^2\right] &= E\left[\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t K_h(x - X_t)\right]^2 \\
&= \frac{1}{T^2}E\left[\sum_{t=1}^{T}\varepsilon_t K_h(x - X_t)\right]^2 \\
&= \frac{1}{T^2}\sum_{t=1}^{T}E[\varepsilon_t^2 K_h^2(x - X_t)] \text{ (by independence, and } E(\varepsilon_t|X_t) = 0) \\
&= \frac{1}{T}E\left[\varepsilon_t^2 K_h^2(x - X_t)\right] \\
&= \frac{1}{T}E\left[\sigma^2(X_t)K_h^2(x - X_t)\right] \text{ (by } E(\varepsilon_t^2|X_t) = \sigma^2(X_t)) \\
&= \frac{1}{T}\int_a^b \left[\frac{1}{h}K\left(\frac{x - y}{h}\right)\right]^2 \sigma^2(y)g(y)dy \\
&= \frac{1}{Th}\sigma^2(x)g(x)\int_{-1}^{1}K^2(u)du[1 + o(1)],
\end{aligned}$$

by change of variable, and the continuity of $\sigma^2(\cdot)g(\cdot)$, where $\sigma^2(x) = E(\varepsilon_t^2|X_t = x)$ is the conditional variance of ε_t or Y_t given $X_t = x$. Note that the variance $E[\hat{V}(x)]^2$ is

proportional to the inverse of Th because Th can be viewed as the effective sample size of the observations falling into the interval $[x-h, x+h]$.

On the other hand, for the denominator, we have as $h \to 0$,

$$\begin{aligned} E\left[\hat{g}(x)\right] &= E\left[K_h(x - X_t)\right] \\ &= \int_a^b \frac{1}{h} K\left(\frac{x-y}{h}\right) g(y) dy \\ &\to g(x) \int_{-1}^{1} K(u) du = g(x) \end{aligned}$$

if $\int_{-1}^{1} K(u)du = 1$. It follows that

$$E\left[\frac{\hat{V}(x)}{E[\hat{g}(x)}\right]^2 = \frac{1}{Th}\frac{\sigma^2(x)}{g(x)} \int_{-1}^{1} K^2(u)du \left[1 + o(1)\right].$$

Thus, the asymptotic variance of $\hat{r}(x)$ is proportional to $(Th)^{-1}$, where Th is the approximate (effective) sample size of the observations in the interval $[x-h, x+h]$. The asymptotic variance of $\hat{r}(x)$ is also proportional to $\sigma^2(x)$ and to $\int_{-1}^{1} K^2(u)du$. Thus, the use of a downward weighting kernel $K(\cdot)$ will reduce the variance of $\hat{r}(x)$ as opposed to the use of the uniform kernel. In other words, it improves the efficiency of the estimator when one discounts observations away from the point x.

For the bias term $\hat{B}(x)$, we first write

$$\hat{B}(x) = E\hat{B}(x) + \left[\hat{B}(x) - E\hat{B}(x)\right].$$

For any given interior point $x \in [a+h, b-h]$, defining $m(x) = r(x)g(x)$, we have

$$\begin{aligned}
E\hat{B}(x) &= E[r(X_t)K_h(x-X_t)] - r(x)E[K_h(x-X_t)] \\
&= \int_a^b r(z)K_h(x-z)g(z)dz - r(x)\int_a^b K_h(x-z)g(z)dz \\
&= \int_a^b m(z)K_h(x-z)dz - r(x)\int_a^b g(z)K_h(x-z)dz \\
&= \int_{(a-x)/h}^{(b-x)/h} m(x+hu)K(u)du \\
&\quad -r(x)\int_{(a-x)/h}^{(b-x)/h} g(x+hu)K(u)du \\
&= m(x)\int_{-1}^{1} K(u)du \\
&\quad +hm'(x)\int_{-1}^{1} uK(u)du \\
&\quad +\frac{1}{2}h^2m''(x)\int_{-1}^{1} u^2K(u)du[1+o(1)] \\
&\quad -r(x)g(x)\int_{-1}^{1} K(u)du \\
&\quad -hr(x)g'(x)\int_{-1}^{1} uK(u)udu \\
&\quad -\frac{1}{2}h^2r(x)g''(x)\int_{-1}^{1} u^2K(u)du][1+o(1)] \\
&= \frac{1}{2}h^2\left[m''(x) - r(x)g''(x)\right]\int_{-1}^{1} u^2K(u)du[1+o(1)] \\
&= \frac{1}{2}h^2[r''(x)g(x) + 2r'(x)g'(x)]C_K + o(h^2),
\end{aligned}$$

where we have used the fact that

$$\begin{aligned}
m''(x) &= [r(x)g(x)]'' \\
&= [r'(x)g(x) + r(x)g'(x)]' \\
&= r''(x)g(x) + 2r'(x)g'(x) + r(x)g''(x)
\end{aligned}$$

It follows that the standardized bias

$$\begin{aligned}
E\left[\frac{\hat{B}(x)}{E\hat{g}(x)}\right] &= \frac{h^2}{2}\left[\frac{m''(x)}{g(x)} - \frac{r(x)g''(x)}{g(x)}\right]C_K + o(h^2) \\
&= \frac{h^2}{2}\left[r''(x) + \frac{2r'(x)g'(x)}{g(x)}\right]C_K + o(h^2),
\end{aligned}$$

where we have made use of the fact that as $h \to 0$,

$$E[\hat{g}(x)] \to g(x) \int_{-1}^{1} K(u)du = g(x)$$

if $\int_{-1}^{1} K(u)du = 1$. Intuitively, the bias of $\hat{r}(x)$ consists of two components: one is $\frac{1}{2}h^2[m''(x)/g(x)]C_K$, which is contributed by the numerator $\hat{m}(x)$; the other is $-\frac{1}{2}h^2[r(x)g''(x)/g(x)]C_K$, which is contributed by the denominator $\hat{g}(x)$, the estimator for density $g(x)$.

Question: Do we have an asymptotically unbiased estimator if $\int_{-1}^{1} K(u)du \neq 1$ (but other conditions on $K(\cdot)$ are the same (i.e., $\int_{-1}^{1} uK(u)du = 0, \int_{-1}^{1} K(u)u^2du = C_K$)?

Yes, we still have

$$E[\hat{B}(x)] = \frac{1}{2}h^2 \left[m''(x) - r(x)g''(x)\right] \int_{-1}^{1} u^2 K(u)du[1 + o(1)].$$

Intuitively, both the numerator and denominator of the kernel regression estimator $\hat{r}(x)$ will produce an integral $\int_{-1}^{1} K(u)du$ and their ratio is unity although the integral itself may not be equal to unity. This ensures that the asymptotic bias is still $O(h^2)$ if $K(u)$ is symmetric about 0. The asymptotic bais will be $O(h)$ if $K(u)$ is not symmetric about 0.

Next, we consider the boundary bias problem for smoothed kernel regression.

Question: What happens to the bias $E[\hat{B}(x)]$ if $x \in [a, a+h) \cup (b-h, b]$, that is, if x is in the boundary regions? In particular, does $E[\hat{B}(x)] \to 0$ as $h \to 0$?

Yes, we still have

$$\frac{E[\hat{B}(x)]}{E[\hat{g}(x)]} = O(h) = o(1)$$

for x in the boundary regions (say, $x = a + \tau h$ for $\tau \in [0, 1]$). This is different from the kernel density estimator $\hat{g}(x)$. This follows because

$$\begin{aligned} E[\hat{B}(x)] &= [m(x) - r(x)g(x)] \int_{-\tau}^{1} K(u)du + O(h) \\ &= O(h) \end{aligned}$$

and therefore

$$\begin{aligned} \frac{E[\hat{B}(x)]}{E[\hat{g}(x)]} &= \frac{[m(x) - r(x)g(x)] \int_{-\tau}^{1} K(u)du}{g(x) \int_{-\tau}^{1} K(u)du} + O(h) \\ &= O(h) \end{aligned}$$

However, the order $O(h)$, which arises due to the fact that $\int_{-\tau}^{1} uK(u)du \neq 0$, is slower than $O(h^2)$, the rate of the bias in the interior region. The boundary correction techniques, such as the use of a jackknife kernel, are useful to further reduce the bias $E[\hat{B}(x)]/E[\hat{g}(x)]$ for x in the boundary regions. They can reduce the bias up to order $O(h^2)$, the same rate as in the interior regions, but still with different proportionalities.

It remains to show that $\hat{B}(x) - E[\hat{B}(x)]$ is a higher order. Again, for simplicity, we first assume that $\{Y_t, X_t\}$ is IID. Put

$$Z_t \equiv Z_t(x) = [r(X_t) - r(x)]\, K_h(x - X_t).$$

Then

$$\begin{aligned}
E[\hat{B}(x) - E\hat{B}(x)]^2 &= E\left[\frac{1}{T}\sum_{t=1}^{T}(Z_t - EZ_t)\right]^2 \\
&= \frac{1}{T^2}\sum_{t=1}^{T} E(Z_t - EZ_t)^2 \text{ by independence} \\
&\leq \frac{1}{T}E(Z_t^2) \\
&= \frac{1}{T}E\left\{[r(X_t) - r(x)]^2 K_h^2(x - X_t)\right\} \\
&\leq \frac{Ch}{T}[1 + o(1)] \text{ (why?)}
\end{aligned}$$

is a higher order term.

It follows that

$$\begin{aligned}
E[\hat{m}(x) - r(x)\hat{g}(x)]^2 &= E[\hat{V}(x) + \hat{B}(x)]^2 \\
&= E[\hat{V}^2(x)] + E[\hat{B}^2(x)] \\
&= E[\hat{V}^2(x)] + E[\hat{B}^2(x)] + E[\hat{B}(x) - E\hat{B}(x)]^2 \\
&= \frac{1}{Th}D_K\sigma^2(x)g(x) \\
&\quad + \frac{h^4}{4}C_K^2[m''(x) - r(x)g''(x)]^2 \\
&\quad + o((Th)^{-1} + h^4).
\end{aligned}$$

Therefore, the asymptotic mean square error (MSE) of $\hat{r}(x)$ is

$$\begin{aligned}
E\left[\hat{r}(x) - r(x)\right]^2 &= \frac{1}{Th}\frac{\sigma^2(x)}{g(x)}D_K + \frac{h^4}{4}\left[\frac{r''(x) + 2r'(x)g'(x)}{g(x)}\right]^2 C_K^2 \\
&\quad + o((Th)^{-1} + h^4) \\
&= O(T^{-1}h^{-1} + h^4).
\end{aligned}$$

The variance is proportional to $(Th)^{-1}$ and the squared bias is proportional to h^4. As a result, an increase in h will reduce the variance but increase the bias, and a decrese in h will increase the variance but reduce the bias. Optimal smoothing can be achieved by balancing the variance and the squared bias. The optimal choice of h is obtained by minimizing the MSE of $\hat{r}(x)$:

$$h^* = c^* T^{-1/5},$$

where the optimal proportionality

$$\begin{aligned} c^* &= \left[\frac{D_K}{C_K^2}\frac{\sigma^2(x)/g(x)}{[m''(x)/g(x) - r(x)g''(x)/g(x)]^2}\right]^{\frac{1}{5}} \\ &= \left[\frac{D_K}{C_K^2}\frac{\sigma^2(x)g(x)}{[r''(x) + 2r'(x)g'(x)]^2}\right]^{\frac{1}{5}} \end{aligned}$$

Thus, the optimal bandwidth h^* should be larger when the data is noisy (i.e., $\sigma^2(x)$ is large) and should be small when the regression function $r(x)$ is not smooth (large derivatives).

Like in density estimation, the optimal kernel for kernel regression estimation remains to be the Epanechnikov kernel

$$K(u) = \frac{3}{4}(1-u^2)\mathbf{1}(|u| < 1).$$

In practice, it is found that the choice of h is more important than the choice of $K(\cdot)$.

Question: How to estimate the derivatives of $r(x)$, such as $r'(x)$ and $r''(x)$ by the kernel method?

One approach is to use $\hat{r}'(x)$ and $\hat{r}''(x)$, assuming that $K(\cdot)$ is twice continuously differentiable. However, it may be noted that the optimal h^* that minimizes the asymptotic MSE of $\hat{r}(\cdot)$ is not the same as the optimal bandwidth h that minimizes the asymptotic MSE of $\hat{r}^{(d)}(\cdot)$, where $d = 1, 2$ respectively. A larger bandwidth is needed to estimate the derivatives of $r(x)$.

In addition to the plug-in method, one can also use the cross-validation method to choose a data-driven bandwidth h. Let $\hat{r}_{(t)}(x)$ be defined in the same way as $\hat{r}(x)$ except the t-th observation (Y_t, X_t) is not used. This is called a "**leave-one-out**" estimator. Then the cross-validation procedure is to choose h to minimize the following sum of

squared residuals, namely,

$$\hat{h}^* = \min_h \sum_{t=1}^{T} \left[Y_t - \hat{r}_{(t)}(X_t)\right]^2$$

The main reason why a leave-one-out estimator is used is that if one used a kernel regression estimator $\hat{r}(X_t)$ using all observations including observation $(Y_t, X_t')'$ at time t, then the optimal bandwidth which minimizes the sum of squared residuals will be obtained by setting $h \to 0$ and $\hat{r}(X_t) = Y_t$. Furthermore, in the IID context, the leave-one-out estimator avoids a nonzero cross-product term which would otherwise "disturb" the mean squared error

$$MSE[\hat{r}(\cdot)] = E\left[Y_t - \hat{r}(X_t)\right]^2.$$

In other words, the sum of squared residuals of the leave-one-out estimator $\hat{r}_{(t)}(X_t)$, scaled by the same size T, will converge to MSE$[\hat{r}(\cdot)]$ plus a constant. As a result, it could be shown (**how?**) that $\hat{h}^*$ will asymptotically minimize MSE$[\hat{r}(\cdot)] = E\left[Y_t - \hat{r}(X_t)\right]^2$. The cross-validation procedure is popular in nonparametric estimation, due to its optimality and robustness (when compared with the plug-in method).

3.2 Local Polynomial Estimation

To provide an motivation for local polynomial smoothing, we now provide an alternative interpretation for the Nadaraya-Watson estimator $\hat{r}(x)$**.** First, we consider a sum of squared residuals (SSR) minimization problem

$$\min_r \sum_{t=1}^{T} (Y_t - r)^2,$$

where r is a constant. The optimal solution is the sample mean

$$\hat{r} = \bar{Y} \equiv \frac{1}{T} \sum_{t=1}^{T} Y_t.$$

Next, we consider a local weighted sum of squared residuals minimization problem

$$\min_r \sum_{t=1}^{T} (Y_t - r)^2 K_h(x - X_t),$$

where r is, again, a real-valued constant. When $K(u)$ has bounded support on $[-1, 1]$, this is the weighted sum of squared residuals to predict the observations $\{Y_t\}$ for which

the values of the corresponding explanatory variables $\{X_t\}$ fall into the interval $[x - h, x + h]$. The FOC is given by

$$\sum_{t=1}^{T}(Y_t - \hat{r})K_h(x - X_t) = 0.$$

It follows that

$$\begin{aligned}\hat{r} &\equiv \hat{r}(x) \\ &= \frac{\sum_{t=1}^{T} Y_t K_h(x - X_t)}{\sum_{t=1}^{T} K_h(x - X_t)} \\ &= \frac{\hat{m}(x)}{\hat{g}(x)}.\end{aligned}$$

This is a local constant estimator. In other words, the Nadaraya-Watson estimator can be viewed as a locally weighted sample mean which mimimizes a locally weighted sum of squared residuals.

Question: Why only use a local constant estimator? Why not use a local linear function? More generally, why not use a local polynomial?

Question: What are the gains, if any, to use a local polynomial estimator?

3.2.1 Local Polynomial Estimator

We now consider a local polynomial estimator for the regression function $r(x)$, where x is a given point in the support of X_t. Suppose z is an arbitrary point in a small neighborhood of x, and $r(z)$ is continuously differentiable with respect to z up to order $p + 1$ in this neighborhood. Then by a $(p + 1)$-order Taylor series expansion, we have for all z in a small neighborhood of a fixed point x,

$$\begin{aligned}r(z) &= \sum_{j=0}^{p} \frac{1}{j!} r^{(j)}(x)(z - x)^j + \frac{1}{(p+1)!} r^{(p+1)}(\bar{x})(z - x)^{p+1} \\ &= \sum_{j=0}^{p} \alpha_j (z - x)^j + \frac{1}{(p+1)!} r^{(p+1)}(\bar{x})(z - x)^{p+1},\end{aligned}$$

where $\bar{x}$ lies in the segment between x and z, and the polynomial coefficient

$$\begin{aligned}\alpha_j &\equiv \alpha_j(x) \\ &= \frac{1}{j!} r^{(j)}(x), \qquad j = 0, 1, ..., p,\end{aligned}$$

depends on the point x. This relation suggests that one can use a local polynomial model to fit the function $r(z)$ in the neighborhood of x as long as the observations in this neighborhood are "sufficiently rich."

Therefore, we consider the following local weighted sum of squared residuals minimization problem

$$\min_{\alpha} \sum_{t=1}^{T} \left[Y_t - \sum_{j=0}^{p} \alpha_j (X_t - x)^j \right]^2 K_h(x - X_t) = \sum_{t=1}^{T} (Y_t - \alpha' Z_t)^2 K_h(x - X_t),$$

where $\alpha = (\alpha_0, \alpha_1, ..., \alpha_p)'$ and $Z_t = Z_t(x) = [1, (X_t - x), ..., (X_t - x)^p]'$. Note that $Z_t = Z_t(x)$ is a $(p+1)$-dimensional local polynomial vector which depends on location x. The resulting local weighted least squares estimator,

$$\hat{r}(z) = \sum_{j=0}^{p} \hat{\alpha}_j (z - x)^j \text{ for all } z \text{ near } x,$$

is the so-called local polynomial estimator of $r(z)$ for z near x, where $\hat{\alpha}$ is the locally weighted least squares estimator, whose expression will be given below.

We now show that the intercept estimator $\hat{\alpha}_0$ is an estimator for $r(x)$, and $\nu!\hat{\alpha}_\nu$ is an estimator for the derivative $r^{(\nu)}(x)$, where $1 \leq \nu \leq p$.

Since

$$\hat{r}(z) = \sum_{j=0}^{p} \hat{\alpha}_j (z - x)^j, \text{ for all } z \text{ near } x,$$

the regression estimator at point x is then given by

$$\hat{r}(x) = \sum_{j=0}^{p} \hat{\alpha}_j (x - x)^j = \hat{\alpha}_0.$$

Moreover, the derivative estimator of $r^{(\nu)}(z)$ for z near the point x is given by

$$\hat{r}^{(\nu)}(z) = \sum_{j=\nu}^{p} j(j-1) \cdots (j - \nu + 1)! \hat{\alpha}_j (z - x)^{j-\nu} \text{ for } \nu \leq p.$$

Thus, we have the ν-th order derivative estimator at point x

$$\hat{r}^{(\nu)}(x) = \nu! \hat{\alpha}_\nu.$$

Interestingly, we can obtain $\hat{r}(x)$ and $\hat{r}^{(\nu)}(x)$ for $1 \leq \nu \leq p$ simultaneously.

Local polynomial smoothing is rather convenient for estimating the $r^{(\nu)}(x)$, $\nu = 0, 1, ..., p$, simultaneously. When $p = 0$, we obtain a local constant estimator, i.e., the

Nadaraya-Watson estimator. Obviously, a local polynomial estimator with $p > 0$ always has a smaller sum of weighted squared residuals than the local constant estimator, because for any local polynomial model, one can always simply set all slope coefficients equal to zero. This implies that the sum of weighted squared residuals will never be larger than that of the local constant estimator.

To compute the local polynomial estimator, one has to choose p, the order of local polynomial, h, the bandwidth, and $K(\cdot)$, the kernel function. Often, a nonnegative kernel function $K(\cdot)$ is used, which corresponds to a second order kernel function. The choices of (p, h) jointly determine the complexity of the local polynomial model. The choice of h is more important than the choice of p (**why?**). It has been recommended that $p = \nu + 1$ if the interest is in estimating $r^{(\nu)}(x)$ for $0 \leq \nu \leq p$. When $p = 1$, it is a local linear smoother. The choice of h can be based on data-driven methods such as the cross-validation or plug-in methods.

To obtain the closed form expression for the local weighted least squares estimator $\hat{\alpha}$, a $(p+1) \times 1$ vector, we put

$$\begin{aligned} Z_t &= Z_t(x) \\ &= [1, (X_t - x), (X_t - x)^2, ..., (X_t - x)^p]', \end{aligned}$$

a $(p+1) \times 1$ polynomial regressor vector, and a weighting function

$$\begin{aligned} W_t &= W_t(x) \\ &= K_h(x - X_t) \\ &= \frac{1}{h} K\left(\frac{x - X_t}{h}\right). \end{aligned}$$

Then the local sum of squared residuals can be written as

$$\begin{aligned} \sum_{t=1}^{T} \left[Y_t - \sum_{j=0}^{p} \alpha_j (X_t - x)^j \right]^2 K_h(x - X_t) &= \sum_{t=1}^{T} (Y_t - \alpha' Z_t)^2 W_t \\ &= (Y - Z\alpha)' W (Y - Z\alpha). \end{aligned}$$

The FOC is

$$\sum_{t=1}^{T} Z_t W_t (Y_t - Z_t' \hat{\alpha}) = 0$$

or equivalently

$$\sum_{t=1}^{T} Z_t W_t Y_t = \left(\sum_{t=1}^{T} Z_t W_t Z_t' \right) \hat{\alpha}.$$

It follows that

$$\begin{aligned}\hat{\alpha} &\equiv \hat{\alpha}(x) \\ &= \left(\sum_{t=1}^{T} Z_t W_t Z_t'\right)^{-1} \sum_{t=1}^{T} Z_t W_t Y_t \\ &= (Z'WZ)^{-1} Z'WY,\end{aligned}$$

where

$$W = W(x) = diag(W_1, \cdots, W_T)$$

is a $T \times T$ diagonal matrix, Z is a $T \times (p+1)$ matrix, and Y is a $T \times 1$ vector. This is a local weighted least squares estimator when $K(\cdot)$ has a bounded support $[-1, 1]$.

Question: What are the advantages of using a local polynomial estimator?

3.2.2 Equivalent Kernel

To exploit the advantages of the local polynomial estimator for $r(x)$, we now investigate its asymptotic properties.

Suppose our interest is in estimating $r^{(\nu)}(x)$, where $0 \leq v \leq p$. Denote $e_{\nu+1}$ for the $(p+1) \times 1$ unit vector with 1 at the $(\nu+1)$-th position and 0 elsewhere. Recalling the weighting function

$$W_t = K_h(x - X_t) = \frac{1}{h} K\left(\frac{x - X_t}{h}\right),$$

we define a j-th order locally weighted sample moment

$$\begin{aligned}\hat{s}_j &= \hat{s}_j(x) \\ &= \sum_{t=1}^{T} (X_t - x)^j K_h(X_t - x) \\ &= \sum_{t=1}^{T} (X_t - x)^j W_t, \quad j = 0, 1, ..., p,\end{aligned}$$

and let

$$\begin{aligned}\hat{S} &= \hat{S}(x) \\ &= Z'WZ \\ &= \sum_{t=1}^{T} Z_t W_t Z_t' \\ &= \left[\hat{s}_{(i-1)+(j-1)}\right]_{(i,j)}\end{aligned}$$

be a $(p+1)\times(p+1)$ stochastic symmetric matrix, whose (i,j)-th element is $\hat{s}_{(i-1)+(j-1)} = \hat{s}_{i+j-2}$.

Then we have the local weighted least squares estimator

$$\hat{\alpha} = \hat{S}^{-1}Z'WY,$$

and so the ν-th component of $\hat{\alpha}$ is given by

$$\begin{aligned}
\hat{\alpha}_\nu &= e'_{\nu+1}\hat{\alpha} \\
&= e'_{\nu+1}\hat{S}^{-1}Z'WY \\
&= e'_{\nu+1}\hat{S}^{-1}\sum_{t=1}^{T} Z_t W_t Y_t \\
&= \sum_{t=1}^{T} e'_{\nu+1}\hat{S}^{-1}\begin{pmatrix} 1 \\ (X_t - x) \\ \dots \\ (X_t - x)^p \end{pmatrix}\frac{1}{h}K\left(\frac{X_t - x}{h}\right)Y_t \\
&= \sum_{t=1}^{T} \hat{W}_\nu\left(\frac{X_t - x}{h}\right)Y_t, \text{ say,}
\end{aligned}$$

where the effective kernel $\hat{W}_\nu(\cdot)$ is the multiplication of the kernel function $K(\cdot)$ with a polynomial function, namely

$$\begin{aligned}
\hat{W}_\nu(u) &= e'_{\nu+1}\hat{S}^{-1}\begin{pmatrix} 1 \\ hu \\ \dots \\ (hu)^p \end{pmatrix}\frac{1}{h}K(u) \\
&= e'_{\nu+1}\hat{S}^{-1}HP(u)\frac{1}{h}K(u),
\end{aligned}$$

where

$$H = diag\{1, h, ..., h^p\}$$

is a $(p+1)\times(p+1)$ diagonal matrix, and

$$P(u) = (1, u, ..., u^p)'$$

is a $(p+1)\times 1$ vector of a p-th order polynomial in u. Note that we will make change of variable $u = (X_t - x)/h$. Obviously, the local polynomial estimator differs from the Nadaraya-Watson estimator in using a different weighting function $\hat{W}_v(\frac{X_t - x}{h})$ for $\{Y_t\}_{t=1}^T$.

Question: What properties does the effective kernel $\hat{W}_\nu(u)$ have?

Lemma [Sample Orthogonality between $\hat{W}_\nu(u)$ and $(X_t - x)^q$]: Let $\hat{W}_v(u)$ be defined as above. Then

$$\sum_{t=1}^{T} \hat{W}_\nu \left(\frac{X_t - x}{h} \right) (X_t - x)^q = \delta_{\nu,q} \text{ for } 0 \leq \nu, q \leq p,$$

where $\delta_{v,q}$ is the Kronecker delta function, namely $\delta_{\nu,q} = 1$ if $\nu = q$ and $\delta_{\nu,q} = 0$ otherwise.

The sample orthogonality between $\hat{W}_\nu(u)$ and $(X_t - x)^q$ is very useful in deriving the bias of the local polynomial estimator $\hat{\alpha}_\nu$.

Question: What is the intuition behind this orthonormality?

Proof: Observing $(X_t - x)^q = Z_t' e_{q+1}$ and $\hat{S} = Z'WZ$, we have

$$\begin{aligned} \sum_{t=1}^{T} \hat{W}_\nu \left(\frac{X_t - x}{h} \right) Z_t' e_{q+1} &= e_{\nu+1}' \hat{S}^{-1} \left(\sum_{t=1}^{T} Z_t W_t Z_t' \right) e_{q+1} \\ &= e_{\nu+1}' I_{p+1} e_{q+1} \\ &= \delta_{\nu q}. \end{aligned}$$

Now, let S be a nonstochastic $(p+1) \times (p+1)$ matrix whose (i,j)th element $S_{(i,j)}$ is $s_{(i-1)+(j-1)} = s_{i+j-2}$, where

$$s_j = \int_{-1}^{1} u^j K(u) du, \quad j = 0, 1, ..., p.$$

Then

$$S = \int_{-1}^{1} P(u) K(u) P(u)' du.$$

Next, we define a nonstochastic equivalent kernel function by

$$\tilde{K}_\nu(u) = e_{\nu+1}' S^{-1} P(u) K(u).$$

This scalar-valued equivalent kernel $\tilde{K}(\cdot)$ has the following properties.

Lemma [Equivalent Kernel]: Suppose $\{Y_t, X_t\}$ is a stationary α-mixing process with $\alpha(j) \leq C_j^{-\beta}$ for $\beta > \frac{5}{2}$, the marginal probability density $g(x)$ of $\{X_t\}$ is bounded on an interval $[a,b]$ and has a continuous derivative at point $x \in [a,b]$, and the kernel function

$K(\cdot)$ satisfies a Lipschitz condition. Then
(1) for any given point x in the interior region $[a+h, b-h]$,

$$\hat{W}_\nu(u) = \frac{1}{Th^{\nu+1}g(x)}\tilde{K}_\nu(u)\left[1 + O_P(a_T)\right],$$

where

$$a_T = [\ln(T)/Th]^{1/2} + h.$$

(2) Moreover, the equivalent kernel $\tilde{K}_\nu(\cdot)$ satisfies the following orthogonality condition

$$\int_{-1}^{1} u^q \tilde{K}_\nu(u)du = \delta_{\nu,q}, \text{ for } 0 \leq \nu, q \leq p.$$

This lemma implies that

$$\begin{aligned}\hat{\alpha}_\nu &= \sum_{t=1}^{T} \hat{W}_\nu\left(\frac{X_t - x}{h}\right) Y_t \\ &= \frac{1}{Th^{\nu+1}g(x)} \sum_{t=1}^{T} \tilde{K}_\nu\left(\frac{X_t - x}{h}\right) Y_t[1 + O_P(a_T)].\end{aligned}$$

In other words, the local polynomial estimator $\hat{\alpha}_v$ works like a kernel regression estimator but with a known probability density $g(x)$ so that there is no need to estimate $g(x)$. This explains why the local polynomial estimator adapts to various design densities, including the boundary region or the region where $g'(x)$ is large in absolute value. Therefore, the bias due to estimation of unknown density $g(x)$ does not enter the MSE criterion of the local polynomial estimator. This is the main advantage of the local polynomial estimator over the Nadaraya-Watson estimator. In particular, it fits well even where $g'(x)$ is large in absolute value. In the regions where $g'(x)$ is large in absolute value, the standard Nadaraya-Watson kernel estimator cannot fit well, due to asymmetric data coverage which yields a large bias. Note that a large value of $g'(x)$ in absolute value implies that the observations $\{X_t\}$ will not be covered symmetrically in a small neighborhood centered at x.

Proof: We first consider the $(p+1) \times (p+1)$ denominator matrix $\hat{S} = [\hat{s}_{(i-1)+(j-1)}]_{(i,j)}$, which is stochastic. Observe that for $j = 0, 1, ..., p$,

$$\frac{1}{Th^j}\hat{s}_j = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{X_t - x}{h}\right)^j K_h(x - X_t)$$

is like a kernel density estimator with the generalized kernel function $u^j K(u)$. Therefore, we have

$$\frac{1}{Th^j}\hat{s}_j = g(x)s_j + O_P(a_T), \quad j = 0, 1, ..., p,$$

where the $O_P(h)$ component in

$$a_T = [(1/Th)^{1/2} \ln T + h]$$

is contributed by the bias term in a first order Taylor series expansion of the integral

$$\begin{aligned} & E\left[\left(\frac{X_t - x}{h}\right)^j K_h(x - X_t)\right] \\ = & \int_a^b \left(\frac{y-x}{h}\right)^j \frac{1}{h} K\left(\frac{x-y}{h}\right) g(y)dy. \end{aligned}$$

Recall $H = diag\{1, h, ..., h^p\}$ and the (i,j)-th element $\hat{S}_{(i,j)} = \hat{s}_{(i-1)+(j-1)}$. It follows that

$$\frac{1}{T}H^{-1}\hat{S}H^{-1} = g(x)S\left[1 + O_P(a_T)\right]$$

or equivalently

$$\hat{S} = Tg(x)HSH\left[1 + O_P(a_T)\right].$$

Substituting this expression into the definition of the effective kernel $\hat{W}_\nu(u)$, we obtain

$$\begin{aligned} \hat{W}_\nu(u) &= e'_{\nu+1}\hat{S}^{-1}HP(u)\frac{1}{h}K(u) \\ &= e'_{\nu+1}\left\{Tg(x)HSH\right\}^{-1}HP(u)\frac{1}{h}K(u)\left[1 + O_P(a_T)\right] \\ &= \frac{1}{Th^{\nu+1}g(x)}\left[e'_{\nu+1}S^{-1}P(u)K(u)\right]\left[1 + O_P(a_T)\right] \\ &= \frac{1}{Th^{\nu+1}g(x)}\tilde{K}_\nu(u)\left[1 + O_P(a_T)\right], \end{aligned}$$

where we have used the fact that $e'_{\nu+1}H = h^\nu e'_{\nu+1}$.

The properties for the equivalent kernel $\tilde{K}_\nu(u)$ can be shown in the same way as the proof of the first part of the lemma for $\hat{W}_\nu(\frac{X_t - x}{h})$. Observing $u^q = P(u)'e_{q+1}$, we have

$$\begin{aligned} \int_{-1}^1 u^q \tilde{K}_\nu(u)du &= \int_{-1}^1 \tilde{K}_\nu(u)u^q du \\ &= e'_{\nu+1}\left[S^{-1}\int_{-1}^1 P(u)K(u)P(u)'du\right]e_{q+1} \\ &= e'_{\nu+1}S^{-1}Se_{q+1} \\ &= e'_{\nu+1}I_{p+1}e_{q+1} \\ &= \delta_{\nu,q}. \end{aligned}$$

This completes the proof of the second part of the lemma for $\tilde{K}_\nu(u)$.

We note that a similar lemma holds for an equivalent boundary kernel function when x is not an interior point in $[a+h, b-h]$. The difference is that the range of the integral has to be changed from $[-1,1]$ to $[-\tau,1]$ or $[-1,\tau]$, depending on whether x is in the left boundary region or the right boundary region.

In other words, the location-dependent weight function $\hat{W}_v[(X_t-x)/h]$ has an equivalent kernel which automatically adapts to the boundary region. See more discussions below.

3.2.3 Asymptotic Properties of Local Polynomial Estimator

Question: What is the MSE of $\hat{\alpha}$?

Noting $Y_t = r(X_t) + \varepsilon_t$, we first write the v-th component of $\hat{\alpha}$,

$$\begin{aligned}
\hat{\alpha}_\nu - \alpha_\nu &= \sum_{t=1}^{T} \hat{W}_\nu\left(\frac{X_t - x}{h}\right) Y_t - \alpha_\nu \\
&= \sum_{t=1}^{T} \hat{W}_\nu\left(\frac{X_t - x}{h}\right) \varepsilon_t + \left[\sum_{t=1}^{T} \hat{W}_\nu\left(\frac{X_t - x}{h}\right) r(X_t) - \alpha_\nu\right] \\
&= \hat{V} + \hat{B}, \text{ say.}
\end{aligned}$$

For the first term $\hat{V}$, using the formula

$$\hat{S} = Tg(x)HSH[1 + O_P(a_T)],$$

which has been proven earlier, we can write

$$\begin{aligned}
\hat{V} &= \sum_{t=1}^{T} \hat{W}_\nu\left(\frac{X_t - x}{h}\right) \varepsilon_t \\
&= e'_{\nu+1}\hat{S}^{-1}Z'W\varepsilon \\
&= \frac{1}{Th^\nu g(x)} e'_{\nu+1}S^{-1}H^{-1}Z'W\varepsilon\left[1 + O_P(a_T)\right],
\end{aligned}$$

where we used the fact that $e'_{\nu+1}H^{-1} = h^{-\nu}e'_{\nu+1}$. Furthermore, assuming that $\{Y_t, X_t\}$

is IID, we have

$$\begin{aligned}
& E(Z'W\varepsilon\varepsilon'WZ) \\
= \; & E\left[\sum_{t=1}^{T}\varepsilon_t Z_t K_h(X_t-x)\right]\left[\sum_{s=1}^{T}\varepsilon_s Z_s' K_h(X_s-x)\right] \\
= \; & \sum_{t=1}^{T} E\left[\varepsilon_t^2 Z_t K_h^2(X_t-x)Z_t'\right] \text{ (by } E(\varepsilon_t|X_t)=0) \\
= \; & TE\left[\varepsilon_t^2 Z_t K_h^2(X_t-x)Z_t'\right] \text{ (by independence)} \\
= \; & \frac{T}{h}\sigma^2(x)g(x)HS^*H[1+o(1)],
\end{aligned}$$

by change of variable and continuity of $\sigma^2(\cdot)$, where S^* is a $(p+1)\times(p+1)$ matrix with (i,j)-th element

$$\begin{aligned}
S^*_{(i,j)} &= s^*_{(i-1)+(j-1)} \\
&= \int_{-1}^{1} u^{(i-1)+(j-1)}K^2(u)du.
\end{aligned}$$

Note that $S^* \neq S$ because the integral here is weighted by $K^2(u)$ rather than $K(u)$.

It follows that the asymptotic variance of $\hat{\alpha}_\nu$,

$$\begin{aligned}
\text{avar}\left(\hat{V}\right) &= \frac{1}{Th^\nu g(x)}e_{\nu+1}'S^{-1}H^{-1}E(Z'W\varepsilon\varepsilon'WZ)H^{-1}S^{-1}\frac{1}{Th^\nu g(x)} \\
&= \frac{1}{Th^{2\nu+1}}\frac{\sigma^2(x)}{g(x)}e_{\nu+1}'S^{-1}S^*S^{-1}e_{\nu+1}[1+o(1)] \\
&= \frac{1}{Th^{2\nu+1}}\frac{\sigma^2(x)}{g(x)}\int_{-1}^{1}\tilde{K}^2(u)du[1+o(1)] \\
&= O(T^{-1}h^{-2\nu-1}),
\end{aligned}$$

where, as before, $\tilde{K}_\nu(u) = e_{\nu+1}'S^{-1}P(u)K(u)$ is the equivalent kernel, and

$$\begin{aligned}
\int_{-1}^{1}\tilde{K}_\nu^2(u)du &= \int_{-1}^{1}[e_{\nu+1}'S^{-1}P(u)K(u)][K(u)P(u)'S^{-1}e_{\nu+1}]du \\
&= e_{\nu+1}'S^{-1}\left[\int_{-1}^{1}P(u)K^2(u)P(u)'du\right]S^{-1}e_{\nu+1} \\
&= e_{\nu+1}'S^{-1}S^*S^{-1}e_{\nu+1}.
\end{aligned}$$

Note that this result is obtained under the IID assumption for $\{Y_t, X_t\}_{t=1}^T$. It still holds under a suitable mixing condition, using an analogous reasoning to that for the kernel density estimator $\hat{g}(x)$.

Question: How to compute the order of magnitude of the bias $\hat{B}$?

Taking a Tailor series expansion around a small neighborhood of x, up to order $p+1$, namely,

$$\begin{aligned} r(X_t) &= \sum_{j=0}^{p} \frac{1}{j!} r^{(j)}(x)(X_t - x)^j + \frac{1}{(p+1)!} r^{(p+1)}(\bar{x}_t)(X_t - x)^{p+1} \\ &= \sum_{j=0}^{p} \frac{1}{j!} r^{(j)}(x)(X_t - x)^j + R(x, X_t), \end{aligned}$$

where $\bar{x}_t$ lies in the segment between x and X_t, we have

$$\begin{aligned} \hat{B} &= \sum_{t=1}^{T} \hat{W}_\nu \left(\frac{X_t - x}{h} \right) r(X_t) - \alpha_\nu \\ &= \sum_{j=0}^{p} \frac{1}{j!} r^{(j)}(x) \sum_{t=1}^{T} \hat{W}_\nu \left(\frac{X_t - x}{h} \right) (X_t - x)^j \\ &\quad - \frac{1}{\nu!} r^{(\nu)}(x) \\ &\quad + \sum_{t=1}^{T} \hat{W}_\nu \left(\frac{X_t - x}{h} \right) R(x, X_t), \end{aligned}$$

where $\alpha_\nu = \frac{1}{\nu!} r^{(\nu)}(x)$, and the reminder

$$R(x, X_t) = \frac{1}{(p+1)!} r^{(p+1)}(\bar{x}_t)(X_t - x)^{p+1},$$

where $\bar{x}_t = \lambda X_t + (1 - \lambda)x$ for some λ in $[0, 1]$. Then using the expression $r(X_t) = \sum_{j=1}^{p} \alpha_j (X_t - x)^j + R(x, X_t)$ and the asymptotic equivalence between $\hat{W}_\nu(\cdot)$ and $\tilde{K}_\nu(\cdot)$, we have

$$\begin{aligned} \hat{B} &= \sum_{t=1}^{T} \hat{W}_\nu \left(\frac{X_t - x}{h} \right) R(x, X_t) \\ &= \frac{1}{T h^{\nu+1} g(x)} \sum_{t=1}^{T} \tilde{K}_\nu \left(\frac{X_t - x}{h} \right) R(x, X_t) \left[1 + O_P(a_T) \right] \\ &= \tilde{B} \left[1 + O_P(a_T) \right], \text{ say,} \end{aligned}$$

by Chebyshev's inequality.

We now consider $\tilde{B}$. It can be shown that

$$\begin{aligned} \tilde{B} - E\tilde{B} &= \frac{1}{T h^{\nu+1} g(x)} \sum_{t=1}^{T} \left\{ \tilde{K}_\nu \left(\frac{X_t - x}{h} \right) R(x, X_t) - E \left[\tilde{K}_\nu \left(\frac{X_t - x}{h} \right) R(x, X_t) \right] \right\} \\ &= O_P(\ln(T)(Th)^{-1/2} h^{-\nu} h^{p+1}) \end{aligned}$$

which is a higher order term. (**Question:** how to show this under the IID assumption and more generally under a suitable mixing condition on $\{Y_t, X_t\}$?). Thus, the bias is determined by

$$\begin{aligned}
E\tilde{B} &= \frac{1}{Th^{\nu+1}g(x)} E\sum_{t=1}^{T} \tilde{K}_\nu\left(\frac{X_t - x}{h}\right) R(x, X_t) \\
&= \frac{1}{Th^{\nu+1}g(x)} E\sum_{t=1}^{T} \tilde{K}_\nu\left(\frac{X_t - x}{h}\right) \frac{r^{(p+1)}(x)}{(p+1)!}(X_t - x)^{p+1} \\
&\quad + \frac{1}{Th^{\nu+1}g(x)} E\sum_{t=1}^{T} \tilde{K}_\nu\left(\frac{X_t - x}{h}\right) \frac{[r^{(p+1)}(\bar{x}_t) - r^{(p+1)}(x)]}{(p+1)!}(X_t - x)^{p+1} \\
&= \frac{h^{p+1}}{h^\nu g(x)} \frac{r^{(p+1)}(x)}{(p+1)!} \int_{-1}^{1} u^{p+1} \tilde{K}_\nu(u)\, g(x+hu) du + O(h^{p+2-\nu}) \\
&= \frac{h^{p+1}}{h^\nu} \frac{1}{(p+1)!} r^{(p+1)}(x) \int_{-1}^{1} u^{p+1} \tilde{K}_\nu(u) du + O(h^{p+2-\nu}) \\
&= \frac{1}{h^\nu} \frac{h^{p+1} r^{(p+1)}(x)}{(p+1)!} e_{\nu+1}' S^{-1} C + O(h^{p+2-\nu}),
\end{aligned}$$

where $C = \int_{-1}^{1} u^{p+1} P(u) K(u) du$ is a $(p+1)\times 1$ vector with the i-th element $\int_{-1}^{1} u^{(p+1)-(i-1)} K(u) du$, and we have made use of the fact that $\tilde{K}_\nu(u) = e_{\nu+1}' S^{-1} P(u) K(u)$.

Question: Why do we have $\int_{-1}^{1} u^{p+1} \tilde{K}_\nu(u) du = e_{\nu+1}' S^{-1} C$?

Recalling $\tilde{K}_\nu(u) = e_{\nu+1}' S^{-1} P(u) K(u)$, we have

$$\begin{aligned}
\int_{-1}^{1} u^{p+1} \tilde{K}_\nu(u) du &= e_{\nu+1}' S^{-1} \int_{-1}^{1} u^{p+1} P(u) K(u) du \\
&= e_{\nu+1}' S^{-1} C.
\end{aligned}$$

It follows that the asymptotic MSE of $\hat{\alpha}_\nu$

$$\begin{aligned}
MSE(\hat{\alpha}_\nu, \alpha_\nu) &= \frac{1}{Th^{2\nu+1}} \frac{\sigma^2(x)}{g(x)} e_{\nu+1}' S^{-1} S^* S^{-1} e_{\nu+1} \\
&\quad + \left[\frac{h^{p+1-\nu} r^{(p+1)}(x)}{(p+1)!}\right]^2 e_{\nu+1}' S^{-1} C C' S^{-1} e_{\nu+1} \\
&= \frac{1}{Th^{2\nu+1}} \frac{\sigma^2(x)}{g(x)} \int_{-1}^{1} \tilde{K}_\nu^2(u) du \\
&\quad + h^{2(p+1-\nu)} \left[\frac{r^{(p+1)}(x)}{(p+1)!}\right]^2 \left[\int_{-1}^{1} u^{p+1} \tilde{K}_\nu(u) du\right]^2 \\
&= O(T^{-1} h^{-2\nu-1} + h^{2(p+1-\nu)}) \\
&= O(T^{-1} h^{-1} + h^4) \text{ if } p = 1, \nu = 0.
\end{aligned}$$

Therefore, the local WLS estimator $\hat{\alpha}_v$ can consistently estimate the Taylor series expansion coefficient α_ν:

$$\nu!\hat{\alpha}_\nu \to^p \nu!\alpha_\nu = r^{(\nu)}(x) \text{ as } T \to \infty.$$

By minimizing the asymptotic MSE, the optimal convergence rate of $\hat{\alpha}_v$ can be achieved by choosing the bandwidth

$$h^* \propto T^{-\frac{1}{2p+3}}.$$

Interestingly, the optimal bandwidth h^* does not depend on the order of the derivative ν. We are using the same h in estimating all $\{\alpha_\nu\}_{\nu=0}^p$. Of course, the proportionality still depends on ν.

The intuitive idea of local polynomial smoothing in econometrics can be dated back to Nerlove (1966), where he uses a piecewise linear regression to estimate a nonlinear cost function for the electricity industry. White (1980, International Economic Review) also has a closely related discussion. He shows that the OLS estimators that are based on the whole sample are not estimating the derivative coefficients in a Taylor series expansion model, unless the regression function $E(Y_t|X_t)$ is a linear function of X_t. White (1980) discusses the case of OLS estimation over the entire support of X_t. Apparently, the OLS estimator can consistently estimate the regression function and its derivaties when a small neighborhood is considered. This is exactly the idea behind local polynomial smoothing.

Next, we use the central limit theory to derive the asymptotic distribution of $\hat{\alpha}_v$.

Theorem [Asymptotic Normality]: If $h = O(T^{-1/(2p+3)})$ and $r^{(p+1)}(x)$ is continuous, then as $T \to \infty$,

$$\sqrt{Th}\left[H(\hat{\alpha} - \alpha) - \frac{h^{p+1}r^{(p+1)}(x)}{(p+1)!}S^{-1}C\right] \to^d N\left(0, \frac{\sigma^2(x)}{g(x)}S^{-1}S^*S^{-1}\right),$$

where $\alpha = [r(x), r^{(1)}(x), ..., r^{(p)}(x)/p!]'$. Therefore,

$$\begin{aligned} & \sqrt{Th^{2\nu+1}}\left[\hat{r}^{(\nu)}(x) - r^{(\nu)}(x) - \frac{h^{p+1-\nu}r^{(p+1)}(x)}{(p+1)!}\int_{-1}^{1} u^{p+1}\tilde{K}_\nu(u)du\right] \\ \to & \ {}^dN\left(0, \frac{(\nu!)^2\sigma^2(x)}{g(x)}\int_{-1}^{1}\tilde{K}_\nu^2(u)du\right). \end{aligned}$$

3.2.4 Boundary Behavior of Local Polynomial Estimator

Question: The above asymptotic results, namely MSE and asymptotic normality, hold for x in the interior region, i.e., $x \in [a+h, b-h]$. What happens if x is in the boundary region?

For simplicity, we assume $[a,b] = [0,1]$ and consider a left boundary point $x = \tau h$ for $\tau \in [0,1]$. Then following reasoning analogous to what we have done for an interior point, we can obtain

$$\begin{aligned}
\text{MSE}[\hat{\alpha}_\nu(\tau h)] &= E\left[\hat{\alpha}_\nu(\tau h) - \alpha_\nu(0)\right]^2 \\
&= \frac{1}{Th^{2\nu+1}} \frac{\sigma^2(0)}{g(0)} e'_{\nu+1} S_\tau^{-1} S_\tau^* S_\tau^{-1} e_{\nu+1} \\
&\quad + \left[\frac{h^{p+1-\nu} r^{(p+1)}(0)}{(p+1)!}\right]^2 e'_{\nu+1} S_\tau^{-1} C_\tau C'_\tau S_\tau^{-1} e_{\nu+1},
\end{aligned}$$

where S_τ, S_τ^* and C_τ are defined in the same way as S, S^* and C, with the lower bounds of all integrals involved being changed from -1 to $-\tau$. For example, S_τ is a $(p+1)\times(p+1)$ matrix, with its (i,j)-th element equal to

$$s_{(i-1)+(j-1),\tau} = \int_{-\tau}^{1} u^{(i-1)+(j-1)} K(u) du.$$

Interestingly, the bias of $\hat{\alpha}_\nu(x)$ is of the same order of magnitude no matter whether x is in the interior region or in the boundary region of $[a,b] = [0,1]$. Of course, the proportionality does depend on the location of x, namely τ if x is in the boundary region. Thus, the local polynomial estimator automatically adapts to the boundary region and does not suffer from the boundary bias problem as the standard kernel method.

Question: What is the intuition behind this? Why does the local polynomial regression estimator behave differently from the Nadaraya-Watson regression estimator? The latter has a bias equal to $O(h)$ in the boundary region.

We consider the local linear estimator (i.e., $p = 1$) as an example. The key here is the joint use of the local intercept and local slope. The latter provides flexibility to adapt to asymmetric data coverage such as those in the boundary regions, as is illustrated in Figure xxx. As a result, the bias of the local linear estimator in the boundary region is much smaller than without using a slope parameter.

An alternative interpretation is that the local polynomial estimator is equivalent to a kernel estimator but with a known density $g(x)$, even when x is in the boundary region. Thus, the boundary bias due to density estimation does not arise for the local polynomial estimator.

We can also obtain an analogous asymptotic normality for $\hat{\alpha}_\nu(\tau h)$ in the boundary region.

Theorem [Asymptotic Normality]: Suppose $h = O(T^{-1/(2p+3)})$ and $r^{(p+1)}(x)$ is continuous. Then as $T \to \infty$,

$$\sqrt{Th}\left[H[\hat{\alpha}(\tau h) - \alpha(0)] - \frac{h^{p+1} r^{(p+1)}(0)}{(p+1)!} S_\tau^{-1} C_\tau\right] \to^d N\left(0, \frac{\sigma^2(0)}{g(0)} S_\tau^{-1} S_\tau^* S_\tau^{-1}\right),$$

where

$$\alpha(0) = [r(0), r^{(1)}(0), ..., r^{(p)}(0)/p!]'.$$

Therefore, as $T \to \infty$,

$$\begin{aligned} & \sqrt{Th^{2\nu+1}}\left[\hat{r}^{(\nu)}(\tau h) - r^{(\nu)}(0) - \frac{h^{p+1-\nu} r^{(p+1)}(0)}{(p+1)!} \int_{-\tau}^{1} u^{p+1} \tilde{K}_{\nu,\tau}(u) du\right] \\ \to \; & {}^d N\left(0, \frac{(\nu!)^2 \sigma^2(0)}{g(0)} \int_{-\tau}^{1} \tilde{K}_{\nu,\tau}^2(u) du\right), \end{aligned}$$

where the equivalent boundary kernel

$$\tilde{K}_{\nu,\tau}(u) = e_{\nu+1}' S_\tau^{-1} P(u) K(u).$$

Proof: The proof is similar to the derivation of the asymptotic MSE for the local polynomial estimator in the interior region.

Question: Why is the local polynomial estimator useful in economic applications?

First of all, it avoids the well-known boundary problem in smoothed kernel regression estimation. Second, it has a smaller bias term for the regression estimator in the areas where the marginal density $g(x)$ of X_t is a steep slope (i.e., when $g'(x)$ is large in absolute value), and consequently is more efficient than the Nadaraya-Watson estimator. A steep slope of $g(x)$ means that the observations will be asymmetrically distributed around x.

Smoothed nonparametric regression estimators have been widely used in econometrics and economics. For example,

- Ait-Sahalia and Lo (1998, *Journal of Finance*) use a multivariate kernel-based regression estimator to estimate the option pricing function

$$\begin{aligned} G_t &= G(X_t, P_t, \tau_t, r_{t,m}) \\ &= \exp[-r_{t,m}(m-t)] \int_{-\infty}^{\infty} Y(p_t, X_t) f_t^*(P_m; T) dp_t, \end{aligned}$$

  where X_t is the strike price at time t, P_t is the price of the underlying asset at time t, m is the length of maturity of the option, and $r_{t,m}$ is the riskfree rate at time t with maturity m.

  They then use

$$\frac{\partial^2 \hat{G}_t}{\partial^2 X_t} = \exp[-r_{t,t}(T-t)] \hat{f}^*(P_t)$$

  to obtain the risk-neutral probability density estimator $\hat{f}^*(P_t)$, which contains rich information about investor preferences and dynamics of data generating process.

- Aït-Sahalia (1996), Stanton (1997), and Chapman and Pearson (1999) use the Nadaraya-Watson estimator $\hat{r}(X_{t-1})$ to estimate $E(X_t|X_{t-1})$, where X_t is the spot interest rate, and examine whether the the drift function $\mu(X_t)$ in the diffusion model

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t,$$

  is linear or nonlinear.

There are many potential topics in time series econometrics that can apply smoothed nonparametric regression estimators. Below are some examples.

Example 1 [Time-Varying CAPM]: Consider a Capital Asset Pricing Model (CAPM) with time-varying parameters:

$$X_{it} = \alpha_i(I_{t-1}) + \beta_i(I_{t-1})'\lambda_t + \varepsilon_{it}, \qquad i = 1, ..., n, t = 1, ..., T,$$

where

$$E(\varepsilon_{it}|I_{t-1}) = 0.$$

This is related to the debate about constant betas versus nonconstant betas. From the Euler equation,

$$\alpha_{it} = \alpha_i(I_{t-1}), \quad \beta_{it} = \beta_i(I_{t-1}),$$

are possibly time-varying coefficients. Suppose $\alpha_{it} = \alpha_i(Z_t)$ and $\beta_{it} = \beta_i(Z_t)$, where Z_t is some state variable or vector in I_{t-1} but the functional forms $\alpha_i(\cdot)$ and $\beta_i(\cdot)$ are unknown. Then one can estimate $\alpha_i(\cdot)$ and $\beta_i(\cdot)$ by solving the local sum of squared residuals minimization problem

$$\min_{\{\alpha_i(\cdot),\beta_i(\cdot)\}} \sum_{i=1}^{n} \sum_{t=1}^{T} [X_{it} - \alpha_i(Z_t) - \beta_i(Z_t)'\lambda_t]^2 K_h(z - Z_t)$$

Question: What is the economic rationale that α_{it} and β_{it} are time-varying?

Kevin Wang (2002, *Journal of Finance*), Ang and Kristense (2012, *Journal of Financial Economics*), and Li and Yang (2012, *Journal of Empirical Finance*) all consider time-varying betas by assuming that the time-varying betas are unknown functions of economic variables or time.

Example 2 [Time-varying Risk Aversion and Equity Risk Premium Puzzle]: Suppose a representative economic agent is solving the lifetime utility maximization problem

$$\max_{\{C_t\}} E\left[\sum_{j=0}^{\infty} \beta^j U(C_{t+j}) \middle| I_t\right]$$

subject to the intertemporal budget constraint

$$C_t = P_t(A_{t+1} - A_t) \leq Y_t + D_t A_t,$$

where C_t is the consumption, A_t is a financial asset, Y_t is the labor income, D_t is the dividends on the asset, and P_t is the price of asset.

The Euler equation for this maximization problem is

$$E\left[\beta\left(\frac{U'(C_{t+1})}{U'(C_t)}\right)\left(\frac{P_{t+1} + D_{t+1}}{P_t}\right) - 1 \middle| I_t\right] = 0,$$

where $(P_{t+1} + D_{t+1})/P_t$ is the gross return on the asset in percentage, $\beta U'(C_{t+1})/U'(C_t)$ is the intertemporal marginal rate of substitution, also called the **stochastic discount factor**. The latter is the time-discounted risk attitude of the representative economic agent.

Suppose the utility function $U(\cdot)$ of the representative economic agent is

$$U(C_t) = \frac{C_t^{1-\gamma} - 1}{1 - \gamma}, \text{ for } \gamma > 0.$$

This is the so-called Constant Relative Risk Aversion (CRRA) utility function. The parameter γ is a measure of the degree of risk aversion. The larger γ, the more risk averse the economic agent.

With the CRRA utility function, the Euler equation becomes

$$E\left[\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}\left(\frac{P_{t+1}+D_{t+1}}{P_t}\right)-1\middle| I_t\right]=0.$$

The unknown parameters β and γ can be estimated using the generalized method of moments. The empirical estimate for γ is too small to justify the observed relatively large difference between stock returns and bond returns. This difficulty is called an "equity risk premium puzzle."

The equity risk premium puzzle exists because the excess of stock returns over returns on investments in bonds is larger than could be explained by standard models of "rational asset" prices. This was first proposed by Mehra and Prescott (1985, "The Equity Premium Puzzle," *Journal of Monetary Economics* 15). Since then, various efforts have been made to explain this puzzle.

Among many other things, a possible solution is to assume both structural parameters β and γ are time-varying, namely $\beta_t=\beta(I_{t-1})$ and $\gamma_t=\gamma(I_{t-1})$, where I_{t-1} is the information set available at time $t-1$. More specifically, we can assume that $\beta_t=\beta(X_t)$ and $\gamma_t=\gamma(X_t)$ for some unknown smooth functions $\beta(\cdot)$ and $\gamma(\cdot)$, where $X_t\in I_{t-1}$ is a state vector that is expected to affect both β_t and γ_t. These time-varying functions can reveal useful information about how the risk attitude of the economic agent changes with the state variables X_t.

Question: How to estimate $\beta(\cdot)$ and $\gamma(\cdot)$?

Recall that the Euler equation is a conditional moment specification, which can be equivalently converted into a generalized regression model:

$$\beta(X_t)\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma(X_t)}\left(\frac{P_{t+1}+D_{t+1}}{P_t}\right)=1+\varepsilon_{t+1},$$

where ε_{t+1} is a stochastic pricing error satisfying the MDS property

$$E(\varepsilon_{t+1}|I_t)=0.$$

As a result, we can estimate the unknown functions $\beta(\cdot)$ and $\gamma(\cdot)$ using minimizing the following local sum of squared generalized residuals:

$$\min_{\beta(\cdot),\gamma(\cdot)} \sum_{t=1}^{T} \left[\beta(X_t) \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma(X_t)} \left(\frac{P_{t+1} + D_{t+1}}{P_t} \right) - 1 \right]^2 K_h \left(\frac{x - X_t}{h} \right)$$

where $\beta(x)$ and $\gamma(x)$ will be estimated by some low-order local polynomial estimators, such as local linear estimators.

More generally, one can incorporate the Euler equations into a time-varying GMM framework

$$E \left\{ Z_t \left[\beta(X_t) \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma(X_t)} \left(\frac{P_{t+1} + D_{t+1}}{P_t} \right) - 1 \right] \right\} = 0,$$

where Z_t is a set of instrumental variables. By approximating $\beta(x)$ and $\gamma(x)$ using a local polynomial respectively, we can obtain local polynomial GMM estimators for $\beta(x)$ and $\gamma(x)$ by minimizing a local quadratic form of the sample moment.

Example 3 [Functional-Coefficient Autoregressive Model]: Suppose $\{X_t\}$ is a strictly stationary time series process and follows a generalized AR(p) process:

$$E(X_t | I_{t-1}) = \sum_{j=1}^{p} \alpha_j(X_{t-d}) X_{t-j},$$

where the autoregressive coefficient $\alpha_j(X_{t-d})$ is a function of X_{t-d}, and the functional form $\alpha_j(\cdot)$ is unknown. The lag order d is called a delay parameter.

Example 4 [Volatility Smile and Derivatives Pricing]: To derive the price of a European call option, Black and Scholes (1973) impose the following assumptions:

- (A1): $d \ln S_t = \mu dt + \sigma dW_t$, where W_t is the Brownian motion, and S_t is the underlying stock price;
- (A2): Frictionless and complete market (no transaction costs; short sales allowed);
- (A3): Constant riskfree interest rate r;
- (A4): European call option, whose payoff function is given by

  $$\phi(S_t) = \max(S_t - K, 0),$$

  where K is the strike price.

Based on a no-arbitrage argument, the following European call option price can be derived:

$$\pi_t = S_0\Phi(d) - Ke^{-rt}\Phi(d - \sigma\sqrt{t}),$$

where $t = m - \tau_t$, m is the maturity period, and

$$d = \frac{\ln(S_0/Ke^{-rt})}{\sigma\sqrt{t}} + \frac{1}{2}\sigma\sqrt{\tau}\ .$$

From the Black-Scholes formula, we can inversely derive the volatility

$$\sigma_t^2 = \sigma^2(K_t, S_t, r_t, \tau_t, \pi_t).$$

This is called the implied volatility. If the pricing formula is correct, then the implied volatility σ_t^2 is a constant function of strike price K_t. This is because σ_t^2 depends only on the data generating process and should not depend on the strike price in any manner. In contrast, if the pricing is incorrect (e.g., the lognormality assumption cannot capture heavy tails of the asset price distribution), then σ_t^2 is generally a convex function of strike price K_t. This is called a volatility smile.

Question: Is the concept of volatility smile well-defined when the distribution of the underlying asset is non-Gaussian (i.e., not lognormal)?

3.2.5 Curse of Dimensionality and Dimension Reduction

Like in multivariate probability density estimation, we will also encounter the curse of dimensionality for regression estimation, when the dimension d of the regressor vector X_t is high. Again, some simplifying assumptions can be made to reduce the curse of dimensionality. Some restrictions on the unknown functions of interest may come from economic theory. For example, a demand function must satisfy the property of a homogeneous function of degree zero. Below are a few examples often used in econometrics and time series econometrics:

Example [Single Index Model]:

$$Y_t = m(X_t'\beta^0) + \varepsilon_t,$$

where $E(\varepsilon_t|X_t) = 0$, the linear combination $X_t'\beta^0$ is a scalar variable, and the functional form $m(\cdot)$ is unknown. Often, the interest is inference of unknown parameter β^0. See Stoker (1986) for more discussion.

Example [Partially Linear Regression Model]:

$$Y_t = X_t'\beta^0 + m(Z_t) + \varepsilon_t,$$

where $E(\varepsilon_t|X_t, Z_t) = 0$, and $m(\cdot)$ is an unknown function of Z_t only. Here, the interest is inference of the marginal effect of the economic variables X_t. However, one has to consistently estimate the unknown function $m(Z_t)$ in order to obtain consistent estimation of parameter β^0.

Example [Functional Coefficient Model]:

$$Y_t = X_t'\beta(Z_t) + \varepsilon_t,$$

where $E(\varepsilon_t|X_t, Z_t) = 0$, and the parameter vector $\beta(\cdot)$ is an unknown function of Z_t only.

Example [Additive Nonlinear Autoregressive Model]:

$$X_t = \sum_{j=1}^{p} \alpha_j(X_{t-j}) + \varepsilon_t,$$

where the $\alpha_j(\cdot)$ functions are unknown.

4 Nonparametric Estimation of Time-Varying Models

In this section, we consider smoothed nonparametric estimation of time-varying models, where model parameters are deterministic functions of time. For simplicity, we focus on estimating a known time trend function in a time series process.

4.1 Slow-Varying Time Trend Estimation

Suppose we observe a bivariate time series random sample $\{Y_t, X_t\}_{t=1}^T$ of size T, where

$$Y_t = m\left(t/T\right) + X_t, \qquad t = 1, ..., T,$$

where $m(\cdot)$ is a smooth but unknown time-trend function and $\{X_t\}$ is a strictly stationary process with $E(X_t) = 0$ and autocovariance function $\gamma(j) = \text{cov}(X_t, X_{t-j})$. Because the mean of Y_t is changing over time, $\{Y_t\}$ is nonstationary.

Question: Why is the trend function $m(\cdot)$ assumed to be a function of normalized time t/T rather than time t only?

This is a crucial device for consistent estimation of the trend function $m(\cdot)$ as the sample size $T \to \infty$. Suppose $m(\cdot)$ is a function of time t only, then when sample size T increases, new information about future times becomes available, but the information about the function $m(\cdot)$ around a given time point, say, t_0, does not increase. Therefore, it is impossible to obtain a consistent estimation of $m(t_0)$. In contrast, with $m(t_0/T)$ as a function of t_0/T, more and more information about $m(\cdot)$ in a neighborhood of t_0/T will become available when T increases, which ensures consistent estimation of $m(t_0/T)$.

Question: How to estimate the time trend function $m(t/T)$?

We can separate the smooth trend from the stochastic component with smoothed nonparametric estimation.

Suppose $m(\cdot)$ is continuously differentiable on $[0,1]$ up to order p, and we are interested in estimating the function $m(t_0/T)$ at a time point t_0 such that $t_0/T \to \tau_0 \in [0,1]$, where τ_0 is a fixed point. Then by a Taylor series expansion, we have, for all t such that t/T lies in a neighborhood of $\tau_0 = t_0/T$,

$$m(t/T) = \sum_{j=0}^{p} \frac{1}{\nu!} m^{(\nu)}(\tau_0) \left(\frac{t-t_0}{T}\right)^j + \frac{1}{(p+1)!} m^{(p+1)}(\bar{\tau}_t) \left(\frac{t-t_0}{T}\right)^{p+1},$$

where $\bar{\tau}_t$ lies in the segment between τ_0 and t/T. Thus, we consider local polynomial smoothing by solving the local weighted sum of squared residuals minimization problem

$$\min_{\alpha} \sum_{t=1}^{T} \left[Y_t - \sum_{j=0}^{p} \alpha_j \left(\frac{t-t_0}{T}\right)^j \right]^2 K_h\left(\frac{t-t_0}{T}\right) = \sum_{t=1}^{T} \left(Y_t - \alpha' Z_t\right)^2 W_t,$$

where $\alpha = (\alpha_0, \alpha_1, ..., \alpha_p)'$,

$$Z_t = \left[1, \left(\frac{t-t_0}{T}\right), ..., \left(\frac{t-t_0}{T}\right)^p\right]',$$

and

$$W_t \equiv K_h\left(\frac{t-t_0}{T}\right) = \frac{1}{h} K\left(\frac{t-t_0}{Th}\right).$$

Then the solution for α is

$$\begin{aligned} \hat{\alpha} &= \left(\sum_{t=1}^{T} Z_t W_t Z_t'\right)^{-1} \sum_{t=1}^{T} Z_t W_t Y_t \\ &= (Z'WZ)^{-1} Z'WY. \end{aligned}$$

In particular, we have

$$\hat{\alpha}_\nu = e'_{\nu+1}\hat{\alpha}, \quad 0 \leq \nu \leq p,$$

where $e_{\nu+1}$ is a $(p+1)\times 1$ unit vector with the $\nu+1$ element being unity and all others being zero.

Question: What are the asymptotic properties of $\hat{\alpha}_\nu$ for $0 \leq \nu \leq p$?

We first derive the asymptotic MSE of $\hat{\alpha}_\nu$.Put

$$\hat{S} = Z'WZ',$$

a nonstochastic $(p+1)\times(p+1)$ matrix, whose (i,j)-element is

$$\hat{S}_{(i,j)} = \sum_{t=1}^{T}\left(\frac{t-t_0}{T}\right)^{(i-1)+(j-1)} K_h\left(t-t_0\right).$$

We first decompose

$$\begin{aligned}
\hat{\alpha}_\nu - \alpha_\nu &= e'_{\nu+1}\hat{S}^{-1}\sum_{t=1}^{T} Z_t W_t X_t \\
&\quad + e'_{\nu+1}\hat{S}^{-1}\sum_{t=1}^{T} Z_t W_t m\left(\frac{t}{T}\right) - \alpha_\nu \\
&= \hat{V} + \hat{B}, \text{ say.}
\end{aligned}$$

For the first term, we have $E(\hat{V}) = 0$ given $E(X_t) = 0$, and

$$\begin{aligned}
\text{var}\left(\hat{V}\right) &= E\left(e'_{\nu+1}\hat{S}^{-1}Z'WXX'WZ\hat{S}^{-1}e_{\nu+1}\right) \\
&= e'_{\nu+1}\hat{S}^{-1}Z'WE\left(X'X\right)WZ\hat{S}^{-1}e_{\nu+1} \\
&= e'_{\nu+1}\hat{S}^{-1}\left[\sum_{t=1}^{T}\sum_{s=1}^{T} Z_t W_t \gamma(t-s) Z'_s W_s\right]\hat{S}^{-1}e_{\nu+1} \\
&= e'_{\nu+1}\hat{S}^{-1}\left[\sum_{j=1-T}^{T-1}\left(1-\frac{|j|}{T}\right)\gamma(j)\sum_{t=1}^{T} Z_t W_t Z'_{t-j} W_{t-j}\right]\hat{S}^{-1}e_{\nu+1}.
\end{aligned}$$

By approximating the discrete sum with a continuous integral, we have

$$\frac{1}{Th}\sum_{t=1}^{T}\left(\frac{t-t_0}{Th}\right)^j K\left(\frac{t-t_0}{Th}\right) \to \int_{-1}^{1} u^j K(u)du \text{ for } 0 \leq j \leq 2p-1$$

as $h \to 0, Th \to \infty$. It follows that

$$\frac{1}{T}H^{-1}\hat{S}H^{-1} = S\left[1+o(1)\right],$$

where, as before, S is a $(p+1)\times(p+1)$ matrix with its (i,j)-th element being $\int_{-1}^{1} u^{(i-1)+(j-1)}K(u)du$. Also, for each given j, by approximating the discrete sum with a continuous integral, we have

$$\frac{1}{Th}\sum_{t=1}^{T}\left(\frac{t-t_0}{Th}\right)^m\left(\frac{t-t_0-j}{Th}\right)^l K\left(\frac{t-t_0}{Th}\right)K\left(\frac{t-t_0-j}{Th}\right)\to\int_{-1}^{1}u^{m+l}K^2(u)du$$

as $h\to 0, Th\to\infty$. Therefore, for any given lah order j, we obtain

$$\frac{1}{T}H^{-1}\left(\sum_{t=1}^{T}Z_tW_tZ'_{t-j}W_{t-j}\right)H^{-1}=h^{-1}S^*[1+o(1)],$$

where S^* is a $(p+1)\times(p+1)$ matrix with its (i,j) element being $\int_{-1}^{1}u^{(i-1)+(j-1)}K^2(u)du$.

It follows that

$$\begin{aligned}\operatorname{var}\left(\hat{V}\right) &= \frac{1}{Th}H^{-1}S^{-1}S^*S^{-1}H^{-1}\left[\sum_{j=-\infty}^{\infty}\gamma(j)\right][1+o(1)]\\ &= \frac{1}{Th^{2\nu+1}}e'_{\nu+1}S^{-1}S^*S^{-1}e_{\nu+1}\left[\sum_{j=-\infty}^{\infty}\gamma(j)\right][1+o(1)]\,.\end{aligned}$$

Unlike the estimator for the regression function $r(X_t)$, the asymptotic variance of the local polynomial estimator $\hat{m}(t_0/T)=\hat{\alpha}_0$ depends on the long-run variance of $\{X_t\}$. In other words, the serial dependence in $\{X_t\}$ has impact on the asymptotic variance of the local polynomial estimator $\hat{\alpha}_0$. More specifically, whether $\{X_t\}$ is IID or serially dependent has an important impact on the asymptotic variance of $\hat{\alpha}_\nu=\frac{1}{\nu!}\hat{m}^{(\nu)}(t_0/T)$.

Question: Why?

The local polynomial estimator for $m(t_0/T)$ is based on the observations in the local interval $[\frac{t_0}{T}-h,\frac{t_0}{T}+h]$. These observations are conservative observations over time in an interval $[t_0-Th,t_0+Th]$, whose width, equal to $2Th$, is increasing but at a slower rate than sample size T. Thus, it maintains the same pattern of serial dependence as the original time series $\{X_t\}$. As a result, the asymptotic variance of $\hat{\alpha}_0$ will depend on the long-run variance of $\{X_t\}$. In contrast, the local polynomial estimator $\hat{r}(x)$ for the regression function $r(x)=E(Y_t|X_t=x)$ is based on the observations in a small interval $[x-h,x+h]$. The observations that fall into this small interval are generally not conservative over time, so their serial dependence structure has been destroyed, and they appear like an IID sequence.

Next, to obtain the bias of $\hat{m}(t_0/T)$, where $t_0 \in [Th, T-Th]$, using the Taylor series expansion,

$$m(t/T) = \sum_{j=0}^{p} \frac{1}{j!} m^{(j)}(\frac{t_0}{T}) \left(\frac{t-t_0}{T}\right)^j + \frac{1}{(p+1)!} m^{(p+1)}(\frac{\bar{t}}{T}) \left(\frac{t-t_0}{T}\right)^{p+1},$$

where $\bar{t} = \lambda t + (1-\lambda)t_0$, we have

$$\begin{aligned}
\hat{B} &= e'_{\nu+1}\hat{S}^{-1} \sum_{t=1}^{T} Z_t W_t m\left(\frac{t}{T}\right) - \alpha_\nu \\
&= \frac{1}{(p+1)!} e'_{\nu+1}\hat{S}^{-1} \sum_{t=1}^{T} Z_t W_h \left(\frac{t-t_0}{T}\right)^{p+1} m^{(p+1)}\left(\frac{\bar{t}}{T}\right) \\
&= \frac{h^{p+1}}{(p+1)!} e'_{\nu+1} H^{-1} \left(H^{-1}\hat{S}H^{-1}\right)^{-1} H^{-1} \sum_{t=1}^{T} Z_t W_h \left(\frac{t-t_0}{Th}\right)^{p+1} m^{(p+1)}\left(\frac{\bar{t}}{T}\right) \\
&= \frac{h^{p+1-\nu} m^{(p+1)}(\frac{t_0}{T})}{(p+1)!} e'_{\nu+1} S^{-1} C \left[1 + o(1)\right],
\end{aligned}$$

where C is a $(p+1)\times 1$ vector with the i-th element being $\int_{-1}^{1} u^{(p+1)-(i-1)} K(u)du$. Here, we have used a continuous integral to approximate a discrete sum:

$$\frac{1}{Th} \sum_{t=1}^{T} \left(\frac{t-t_0}{Th}\right)^j K\left(\frac{t-t_0}{Th}\right) \to \int_{-1}^{1} u^j K(u)du$$

as $T \to \infty$. It follows that the asymptotic MSE of $\hat{\alpha}_\nu$ is

$$\begin{aligned}
MSE(\hat{\alpha}_\nu) &= \frac{1}{Th^{2\nu+1}} e'_{\nu+1} S^{-1} S^* S^{-1} e_{\nu+1} \sum_{j=-\infty}^{\infty} \gamma(j) \\
&\quad + h^{2(p+1-\nu)} \left[\frac{m^{(p+1)}\left(\frac{t_0}{T}\right)}{(p+1)!}\right]^2 \left(e'_{\nu+1} S^{-1} C\right)^2 \\
&\quad + o(T^{-1}h^{-2\nu-1} + h^{2(p+1-\nu)}).
\end{aligned}$$

Next, we derive the asymptotic distribution of $\hat{m}(t_0/T)$.

Next, we use the central limit theorem to derive the asymptotic distribution of $\hat{m}(t_0/T)$.

Theorem [Asymptotic Normality of $\hat{m}(t_0/T)$]: If $h = O(T^{-1/(2p+3)})$ and $m^{(p+1)}(t_0/T)$ is continuous, where $t_0 \in [Th, T-Th]$, then as $T \to \infty$,

$$\begin{aligned}
&\sqrt{Th}\left[H[\hat{\alpha}(t_0/T) - \alpha(t_0/T)] - \frac{h^{p+1} m^{(p+1)}(t_0/T)}{(p+1)!} S^{-1} C\right] \\
\to\ & {}^d N\left(0, S^{-1} S^* S^{-1} \sum_{j=-\infty}^{\infty} \gamma(j)\right),
\end{aligned}$$

where

$$\alpha(0) = [r(0), r^{(1)}(0), ..., r^{(p)}(0)/p!]'.$$

Therefore,

$$\sqrt{Th^{2\nu+1}}\left[\hat{m}^{(\nu)}(t_0/T) - m^{(\nu)}(t_0/T) - \frac{h^{p+1-\nu}m^{(p+1)}(t_0/T)}{(p+1)!}\int_{-1}^{1} u^{p+1}K_\nu^*(u)du\right]$$
$$\rightarrow \ {}^dN\left(0, (\nu!)^2\int_{-1}^{1}\tilde{K}^2(u)du\sum_{j=-\infty}^{\infty}\gamma(j)\right).$$

Similar results for the asymptotic MSE and asymptotic normality of the local polynomial trend estimator for t_0 in the boundary region $[1, Th)$ or $(T - Th, T]$ could also be obtained. We omit them for space.

4.2 Locally Linear Time-Varying Regression Estimation

Question: How to model smooth time changes in economics?

We consider a locally linear time-varying regression model

$$Y_t = X_t'\alpha\left(\frac{t}{T}\right) + \varepsilon_t, \qquad t = 1, ..., T,$$

where Y_t is a scalar, X_t is a $d \times 1$ random vector, and $\alpha(t/T)$ is a $d \times 1$ smooth function of normalized time t/T. The time-trend time series model can be viewed as a special case of this locally linear time-varying regression model with $X_t = 1$.

Question: Why to consider smooth changes in a regression model?

Structural changes are rather a rule than an exception, due to advances in technology, changes in preferences, policy shifts, and institutional changes in the economic system.

It takes time for economic agents to react to sudden shocks, because it takes time for economic agents to collect information needed for making decisions, it takes time for markets to reach some consensus due to heterogenous beliefs, it takes time for economic agents to change their habits, etc. Even if individual agents can respond immediately to sudden changes, the aggregated economic variables (such as consumption) over many individuals will become smooth. Indeed, as Alfred Marshall points out, economic changes are evolutionary.

The locally linear time-varying regression model is potentially useful for macroeconomic applications and for long time series data.

Question: How to estimate the $d \times 1$ time-varying parameter vector $\alpha(t/T)$?

Suppose $\alpha(t/T)$ is continuously differentiable up to order $p+1$, and t_0 is a specified time point such that t_0/T converges τ_0 in $[0,1]$. Then by a Taylor series expansion, we have that for all t such that t/T is in a small neighborhoo of t_0/T,

$$\begin{aligned}\alpha\left(\frac{t}{T}\right) &= \sum_{j=0}^{p}\frac{1}{j!}\alpha^{(j)}\left(\frac{t_0}{T}\right)\left(\frac{t-t_0}{T}\right)^j \\ &\quad + \frac{1}{(p+1)!}\alpha^{(p+1)}\left(\frac{\bar{t}}{T}\right)\left(\frac{t-t_0}{T}\right)^{p+1},\end{aligned}$$

where $\bar{t}$ lies in the segment between t and t_0. Precisely, $\alpha^{(p+1)}\left(\bar{t}/T\right)$ should be understood as being evaluated at a different $\bar{t}$ for a different component of the $d \times 1$ vector $\alpha^{(p+1)}(\cdot)$. Therefore, we can use a p-th order polynomial to approximate the unknown vector function $\alpha(t/T)$ in the neighbood of t_0/T. Put

$$Z_t = \left[1, \frac{t-t_0}{T}, ..., \left(\frac{t-t_0}{T}\right)^p\right]'$$

and

$$Q_t = Z_t \otimes X_t$$

is a $d(p+1) \times 1$ vector of augmented regressors, where $\otimes$ is the Konecker product. Then we consider the following locally weighted sum of squared residuals minimization problem

$$\begin{aligned}&\sum_{t=1}^{T}\left(Y_t - \sum_{l=1}^{d}\alpha_{lt}'X_{lt}\right)^2 K_h(t-t_0) \\ =\ &\sum_{t=1}^{T}\left(Y_t - \sum_{l=1}^{d}\alpha_l'Z_tX_{lt}\right)^2 K_h(t-t_0) \\ =\ &\sum_{t=1}^{T}\left[Y_t - \alpha'(Z_t \otimes X_t)\right]^2 K_h(t-t_0) \\ =\ &\sum_{t=1}^{T}\left(Y_t - \alpha'Q_t\right)^2 K_h(t-t_0),\end{aligned}$$

where $\alpha = (\alpha_1', ..., \alpha_d')'$ is a $d(p+1) \times 1$ vector, with α_l being a $(p+1) \times 1$ coefficient vector for the product regressor vector Z_tX_{lt}.

The local polynomial estimator

$$\begin{aligned}\hat{\alpha} &= \left(\sum_{t=1}^{T} Q_t W_t Q_t'\right)^{-1} \sum_{t=1}^{T} Q_t W_t Y_t \\ &= \left[\sum_{t=1}^{T} (Z_t \otimes X_t) W_t (Z_t \otimes X_t)'\right]^{-1} \sum_{t=1}^{T} (Z_t \otimes X_t) W_t Y_t.\end{aligned}$$

The estimator for the $d \times 1$ vector $\alpha(t_0/T)$ is then given by

$$\hat{\alpha}(t_0/T) = (I_d \otimes e_1)'\hat{\alpha},$$

where I_d is a $d \times d$ identity matrix and e_1 is a $(p+1) \times 1$ vector with unity for the first component equal to 1 and all the other components equal to zero. Intuitively, $\hat{\alpha}(t_0/T)$ is a $d \times 1$ vector consisting of the d estimated intercepts from $\{\hat{\alpha}_l\}_{l=1}^{d}$, where $\hat{\alpha}_l$ is a $(p+1) \times 1$ estimated coefficient vector for $Z_t X_{lt}$.

We could derive the asymptotic MSE formula for $\hat{\alpha}(t_0/T)$, and their asymptotic normal distributions.

Theorem [Asymptotic MSE]: Suppose $\{X_t', \varepsilon_t\}'$ is a strictly stationary α-mixing process with $\gamma(j) = \text{cov}(\varepsilon_t, \varepsilon_{t-j})$ and $Q = E(X_t X_t')$, and $\{X_t\}$ and $\{\varepsilon_t\}$ are mutually independent. Then for any given t_0/T in the interior region of $[0,1]$, the asymptotic MSE of each component of the $d \times 1$ vector $\hat{\alpha}(t_0/T) = (I_d \otimes e_1)'\hat{\alpha}$, where $\hat{\alpha} = (\hat{\alpha}_1', ..., \hat{\alpha}_d')'$ is the local weighted least squares estimator, is given by

$$\begin{aligned}MSE\left[\hat{\alpha}_l\left(t_0/T\right), \alpha_l(t_0/T)\right] &= \frac{1}{Th}\left[\sum_{j=-\infty}^{\infty} \gamma(j)\right] (I_d \otimes e_1)' (Q \otimes S)^{-1} (Q \otimes S^*) (Q \otimes S)^{-1} (I_d \otimes e_1) \\ &\quad + \left[\frac{h^{p+1}\alpha_l^{(p+1)}(t_0/T)}{(p+1)!}\right]^2 (I_d \otimes e_1)' (Q \otimes S)^{-1} (Q \otimes C) (Q \otimes C)' (Q \otimes S)^{-1} \\ &\quad + o(T^{-1}h^{-1} + h^{2(p+1)}) \\ &= \frac{1}{Th}\left[\sum_{j=-\infty}^{\infty} \gamma(j)\right] \int_{-1}^{1} \tilde{K}_0^2(u)du \\ &\quad + h^{2(p+1)}\left[\frac{\alpha_l^{(p+1)}(t_0/T)}{(p+1)!}\right]^2 \left[\int_{-1}^{1} u^{p+1}\tilde{K}_0(u)du\right]^2 \\ &\quad + o(T^{-1}h^{-1} + h^{2(p+1)}),\end{aligned}$$

where the equivalent kernel is defined as

$$\tilde{K}_0(u) = (I_d \otimes e_1)' (Q \otimes S)^{-1} [Q \otimes P(u)] K(u).$$

By minimizing the asymptotic MSE, the optimal convergence rate of each component of $\hat{\alpha}(t_0/T)$ can be obtained by choosing the bandwidth

$$h^* \propto T^{-\frac{1}{2p+3}}.$$

Theorem [Asymptotic Normality of $\hat{\alpha}(t_0/T)$]: If $h = O(T^{-1/(2p+3)})$ and $\alpha^{(p+1)}(t_0/T)$ is continuous, where $t_0 \in [Th, T - Th]$, then for $l = 1, ..., d$, as $T \to \infty$,

$$\begin{aligned} & \sqrt{Th} \left[\hat{\alpha}_l(t_0/T) - \alpha_l(t_0/T)] - \frac{h^{p+1} \alpha_l^{(p+1)}(t_0/T)}{(p+1)!} (Q \otimes S)^{-1} (Q \otimes C) \right] \\ \to \; & {}^d N \left(0, (Q \otimes S)^{-1} (Q \otimes S^*) (Q \otimes S)^{-1} \sum_{j=-\infty}^{\infty} \gamma(j) \right). \end{aligned}$$

Similar results of the asymptotic MSE and asymptotic normality for $\hat{\alpha}(t_0/T)$ can be obtained when the normalized time time $\frac{t_0}{T}$ is in the left boundary region $[0, h]$ or the right boundary region $[1 - h, 1]$.

By plotting the estimator $\hat{\alpha}(t_0/T)$ as a function of t_0/T in the interval $[0, 1]$, we can examine whether the coefficient vector α is time-varying. Formally, Chen and Hong (2012) propose some consistent tests for smooth structural changes as well as a finite number of multiple breaks in a linear regression model by using a local linear estimator for the time-varying coefficient $\alpha(\cdot)$. Specifically, they propose a generalized Chow (1960) test, which is a generalized F-test by comparing the sum of squared residuals of a local linear regression model with that of a constant parameter regression model. They also propose a generalized Hausman (1978) type test by comparing the fitted values of a local linear estimator with those of a constant parameter regression model. Chen and Hong (2012) derive the asymptotic null distribution of these test statistics under the null hypothesis of no structural change and establish consistency of these tests under the alternative hypothesis. See Chen and Hong (2012) for more discussion, including a simulation study and an empirical application.

Chen and Hong (2016) also propose a local smoothing quasi-likelihood based test for parameter constancy of a GARCH model.

5 Nonparametric Estimation in Frequency Domain

Questions: Given a time series random sample $\{X_t\}_{t=1}^T$ of size T,

- How to estimate the power spectral density $h(\omega)$ of $\{X_t\}$?
- How to estimate the bispectral density $b(\omega_1, \omega_2)$ of $\{X_t\}$?
- How to estimate the generalized spectral density $f(\omega, u, v)$ of $\{X_t\}$?

5.1 Sample Periodogram

Suppose $\{X_t\}$ is a weakly stationary time series with autocovariance function $\gamma(j)$ and the spectral density function $h(\omega)$. For simplicity, we assume $E(X_t) = 0$ and we know it. Then we can estimate $\gamma(j)$ the sample autocovariance function

$$\hat{\gamma}(j) = T^{-1} \sum_{t=|j|+1}^{T} X_t X_{t-|j|}, \qquad j = 0, \pm 1, ..., \pm(T-1).$$

If μ is unknown, we should use the sample autocovariance function

$$\hat{\gamma}(j) = T^{-1} \sum_{t=|j|+1}^{t} (X_t - \bar{X})(X_{t-|j|} - \bar{X}),$$

where $\bar{X}$ is the sample mean. The asymptotic analysis is a bit more tedious but the same results can be obtained since the replacement of μ by $\bar{X}$ has no impact on the asymptotic results below.

In this section, our interest is in consistent estimation of the spectral density function $h(\omega)$ based on an observed random sample $\{X_t\}_{t=1}^T$. Recall the spectral density function is

$$h(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma(j) e^{-ij\omega}.$$

For a WN$(0, \sigma^2)$ process, the spectral density

$$h(\omega) = \frac{1}{2\pi} \gamma(0), \quad \omega \in [-\pi, \pi],$$

where $\gamma(0) = \text{var}(X_t)$. In this case, a spectral density estimator is

$$\hat{h}(\omega) = \frac{1}{2\pi}\hat{\gamma}(0).$$

For an MA(1) process, the spectral density

$$h(\omega) = \frac{1}{2\pi}\gamma(0) + \frac{1}{\pi}\gamma(1)\cos(\omega).$$

The corresponding spectral estimator is

$$\hat{h}(\omega) = \frac{1}{2\pi}\hat{\gamma}(0) + \frac{1}{\pi}\hat{\gamma}(1)\cos(\omega).$$

For an ARMA(p, q) process, the spectral density function

$$h(\omega) = \frac{\sigma^2}{2\pi}\left|\frac{1+\sum_{j=1}^{q}\theta_j e^{-ij\omega}}{1-\sum_{j=1}^{p}\phi_j e^{-ij\omega}}\right|^2,$$

where $\sigma^2 = E(\varepsilon_t^2)$ is the variance of innovation ε_t.

A spectral density estimator is

$$\hat{h}(\omega) = \frac{\hat{\sigma}^2}{2\pi}\left|\frac{1+\sum_{j=1}^{q}\hat{\theta}_j e^{-ij\omega}}{1-\sum_{j=1}^{p}\hat{\phi}_j e^{-ij\omega}}\right|^2$$

where $(\hat{\theta}_j, \hat{\phi}_j)$ are consistent parameter estimators, and

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=\max(p,q)+1}^{T}\hat{\varepsilon}_t^2,$$

where

$$\hat{\varepsilon}_t = X_t - \sum_{j=1}^{p}\hat{\phi}_j X_{t-j} - \sum_{j=1}^{q}\hat{\theta}_j\hat{\varepsilon}_{t-j},$$

with the initial values $\hat{\varepsilon}_t = 0$ for all $t \leq 0$.

For a general linear process (or when we do not know what process X_t is), we may like to use the sample periodogram as a spectral density estimator:

$$\begin{aligned}
\hat{I}(\omega) &= \frac{1}{2\pi T}\left|\sum_{t=1}^{T}X_t e^{it\omega}\right|^2 \\
&= \frac{1}{2\pi}\sum_{j=1-T}^{T-1}\left(1-\frac{|j|}{T}\right)\hat{\gamma}(j)e^{-ij\omega} \\
&= \frac{1}{2\pi}\hat{\gamma}(0) + \frac{1}{\pi}\sum_{j=1}^{T-1}\left(1-\frac{j}{T}\right)\hat{\gamma}(j)\cos(j\omega).
\end{aligned}$$

The sample periodogram $\hat{I}(\omega)$, based on the time series random sample $\{X_t\}_{t=1}^T$, is the squared modulus of the discrete Fourier transform of $\{X_t\}_{t=1}^T$.

Unfortunately, this sample periodogram $\hat{I}(\omega)$ is not consistent for the spectral density $h(\omega)$. Why?

To explain, let us consider the simplest case when $\{X_t\}$ is IID. Then we have $h(\omega) = \frac{1}{2\pi}\gamma(0)$, and

$$E\hat{I}(\omega) = \frac{1}{2\pi}\gamma(0) = h(\omega)$$

so the bias $E[\hat{I}(\omega)] - h(\omega) = 0$ for all $\omega \in [-\pi, \pi]$.
On the other hand, under the IID condition, we have

$$\text{cov}[\sqrt{T}\hat{\gamma}(i), \sqrt{T}\hat{\gamma}(j)] = \begin{cases} (1 - |i|/T)\gamma^2(0) & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

It follows that

$$\begin{aligned} \text{var}[\hat{I}(\omega)] &= \frac{1}{(2\pi)^2}\text{var}[\hat{\gamma}(0)] \\ &\quad + \frac{1}{(\pi)^2}\sum_{j=1}^{T-1}\left(1 - \frac{j}{T}\right)^2 \text{var}[\hat{\gamma}(j)]\cos^2(j\omega) \\ &= C_0\frac{1}{T} + C_1\frac{1}{T}\sum_{j=1}^{T-1}\left(1 - \frac{j}{T}\right)^2 \cos^2(j\omega) \\ &\leq C_0\frac{1}{T} + C_1 \cdot \frac{1}{2} \\ &= O(1). \end{aligned}$$

The variance $\text{var}[\hat{I}(\omega)]$ never decays to 0 as $T \to \infty$, and so $\hat{I}(\omega)$ is not consistent for $h(\omega)$.

Why? There are too many estimated coefficients $\{\hat{\gamma}(j)\}_{j=0}^{T-1}$! There is a total of T estimated coefficients, where T is the sample size.

We now offer an alternative explanation why the sample periodogram $\hat{I}(\omega)$ is not

consistent for $h(\omega)$. Consider the Integrated MSE (IMSE) of $\hat{I}(\omega)$,

$$\begin{aligned}
& \text{IMSE}(\hat{I}) \\
= \; & E\int_{-\pi}^{\pi}\left|\hat{I}(\omega)-h(\omega)\right|^2 d\omega \\
= \; & E\int_{-\pi}^{\pi}\left|\hat{I}(\omega)-E\hat{I}(\omega)\right|^2 d\omega \\
& + \int_{-\pi}^{\pi}\left|E\hat{I}(\omega)-h(\omega)\right|^2 d\omega \\
= \; & E\left[\frac{1}{2\pi}\sum_{j=1-T}^{T-1}\left(1-\frac{|j|}{T}\right)^2[\hat{\gamma}(j)-E\hat{\gamma}(j)]^2\right] \\
& + \left[\frac{1}{2\pi}\sum_{|j|<T}\left[\left(1-\frac{|j|}{T}\right)E\hat{\gamma}(j)-\gamma(j)\right]^2+\frac{1}{2\pi}\sum_{|j|\geq T}\gamma^2(j)\right],
\end{aligned}$$

by orthogonality of exponential bases $\{e^{\mathbf{i}j\omega}\}$, or the so-called Parseval's identity.

Note that given $E(X_t)=0$,

$$\begin{aligned}
E\hat{\gamma}(j) &= T^{-1}\sum_{t=|j|+1}^{T}E(X_tX_{t-|j|}) \\
&= (1-|j|/T)\gamma(j),
\end{aligned}$$

so we have the squared bias

$$\sum_{|j|<T}[E\hat{\gamma}(j)-\gamma(j)]^2=\sum_{|j|<T}(j/T)^2\gamma^2(j)\to 0$$

if $\sum_{j=-\infty}^{\infty}\gamma^2(j)<\infty$. For the last term, we also have $\sum_{|j|>T}\gamma^2(j)\to 0$ as $T\to\infty$.

For the first term of IMSE$(\hat{I})$, we have

$$\begin{aligned}
\sum_{j=1-T}^{T-1}E[\hat{\gamma}(j)-E\hat{\gamma}(j)]^2 &= \sum_{j=1-T}^{T-1}\text{var}[\hat{\gamma}(j)] \\
&= O(1).
\end{aligned}$$

because $\text{var}[\hat{\gamma}(j)]=CT^{-1}$ as $T\to\infty$ under certain regularity conditions (e.g., a mixing condition with a suitable rate), for some $C>0$. Therefore, the variance of the periodogram $\hat{I}(\omega)$ does not vanish as $T\to\infty$.

5.2 Kernel Spectral Estimation

Question: What is a solution to the inconsistency of the sample periodogram $\hat{I}(\omega)$?

In response to the fact that the sample periodogram $\hat{I}(\omega)$ is not a consistent estimator for the spectral density $h(\omega)$ because it contains "too many" estimated parameters, one can consider a truncated spectral density estimator,

$$\hat{h}(\omega) = \frac{1}{2\pi} \sum_{j=-p}^{p} \hat{\gamma}(j) e^{-ij\omega}, \qquad \omega \in [-\pi, \pi],$$

where p is the maximum truncation lag order such that $p = p(T) \to \infty, p/T \to 0$ as $T \to \infty$. Thus, the number $p+1$ of the estimated parameters $\{\hat{\gamma}(j)\}_{j=0}^{p}$ is substantially smaller than the sample T when $T \to \infty$. As a result, the variance of $\hat{h}(\omega)$ is expected to vanish to zero as $T \to \infty$.

The truncated spectral density estimator was used by Hansen (1980) and White and Domowitz (1984) to consistently estimate the asymptotic variance-covariance matrix of econometric estimators (e.g., GMM, OLS) in time series contexts, which is proportional to the spectral density of certain time series process at frequency zero (see Chapter 3). However, such an estimator may not be positive semi-definite in finite samples. This may cause some trouble in applications.

To ensure a positive semi-definite variance-covariance matrix estimator, we can use a weighted estimator

$$\hat{h}(\omega) = \frac{1}{2\pi} \sum_{j=-p}^{p} k(j/p) \hat{\gamma}(j) e^{-ij\omega},$$

where $k(\cdot)$ is a kernel function for the lag order, and so it is also called a lag window. An example is the Bartlett kernel ,

$$k(z) = (1 - |z|)\mathbf{1}(|z| \leq 1),$$

where $\mathbf{1}(\bullet)$ is the indicator function. This is used in Newey and West (1987, 1994). The Bartlett kernel-based spectral density estimator at frequency zero is always positive semi-definite (**why?**). We note that the sample periodogram $\hat{I}(\omega)$ can be viewed as a spectral density estimator based on the Bartlett kernel with the choice of the maximum truncation lag order $p = T$.

Question: What is the advantage of introducing the kernel function $k(\cdot)$?

Most kernel functions are downward weighting, i.e., they discount higher order lags. As a result, they help reduce the variance of $\hat{h}(\omega)$.

For a weakly stationary process with square-summable autocovariances, serial correlation decays to zero as lag j increases. This is consistent with the stylized fact that the remote past events have smaller impacts on the current economic systems and financial markets than the recent events. Given this, it makes sense to discount higher order lags, namely to discount remote past events.

More generally, we can consider

$$\hat{h}(\omega) = \frac{1}{2\pi} \sum_{j=1-T}^{T-1} k(j/p)\hat{\gamma}(j)e^{-ij\omega}, \qquad \omega \in [-\pi, \pi],$$

where $k(\cdot)$ is allowed to have unbounded support, so that all $T-1$ sample autocovariances are used in spectral estimation. An example is the Daniell kernel

$$k(z) = \frac{\sin(\pi z)}{\pi z}, \qquad z \in \mathbb{R}.$$

As will be seen below, the optimal kernel that minimizes the MSE of the kernel spectral density estimator $\hat{h}(\omega)$ also has an unbounded support (see the Quadratic-Spectral kernel below).

When $k(\cdot)$ has an unbounded support, p can no longer be viewed as a maximum lag truncation order but a smoothing parameter.

We impose the following regularity condition on the kernel function.

Assumption [Kernel Function]: $k(\cdot)$ is a symmetric function that is continuous at all but a finite number of points, such that (i) $|k(z)| \leq 1$, (ii) $k(0) = 1$,(iii) $\int_{-\infty}^{\infty} k^2(z)dz < \infty$, and (iv) there exists a positive real number q such that

$$0 < k_q = \lim_{z \to 0} \frac{k(0) - k(z)}{|z|^q} < \infty.$$

The quantity k_q characterizes the speed at which $k(z)$ converges to $k(0)$ in the neighborhood of 0. For the Bartlett kernel, $q = 1$. For the Daniell kernel, $q = 2$.

Question: What is the relationship between the kernel function or lag window $k(\cdot)$ and the kernel function $K(\cdot)$ used for density and regression estimation?

In terms of mathematical properties, the kernel function or lag window $k(\cdot)$ used for spectral density estimation is the Fourier transform of a kernel function $K(\cdot)$ used in

probability density and regression estimation, namely

$$K(u) = \frac{1}{2\pi}\int_{-\infty}^{\infty} k(z)e^{-\mathbf{i}zu}dz,$$

and

$$k(z) = \int_{-\infty}^{\infty} K(u)e^{\mathbf{i}uz}du.$$

When $K(u)$ is a positive kernel (i.e., a symmetric probability density function), $k(z)$ is the characteristic function of the probability distribution $K(u)$. Thus, it is straightforward to understand the conditions imposed on $k(z)$ and $K(u)$ are essentially the same:

- **[Unity Integral]**: $k(0) = 1$ is equivalent to $\int_{-\infty}^{\infty} K(u)du = 1$;
- **[Square-Integratability]**: $\int_{-\infty}^{\infty} k^2(z)dz = 2\pi\int_{-\infty}^{\infty} K^2(u)du < \infty$;
- **[Finite Variance]**: For $q = 2$,

$$k_2 = -\frac{1}{2}k''(0) = \frac{1}{2}\int_{-\infty}^{\infty} u^2K(u)du < \infty.$$

- **[Symmetry]**: The symmetry of $k(z)$ is equivalent to the symmetry of $K(u)$, so $\int_{-\infty}^{\infty} uK(u)du = 0$.

We now provide some commonly used kernels in practice. They include:

- Truncated kernel

$$k(z) = \mathbf{1}(|z| \leq 1).$$

  Its Fourier transform

$$K(u) = \frac{1}{\pi}\frac{\sin u}{u}, \quad -\infty < u < \infty.$$

- Bartlett kernel

$$k(z) = (1 - |z|)\mathbf{1}(|z| \leq 1).$$

  Its Fourier transform

$$K(u) = \frac{1}{2\pi}\left[\frac{\sin(u/2)}{u/2}\right]^2, \quad -\infty < u < \infty.$$

- Daniell kernel

$$k(z) = \frac{\sin(\pi z)}{\pi z}, \quad -\infty < z < \infty.$$

Its Fourier transform

$$K(u) = \frac{1}{2\pi}\mathbf{1}(|u| \leq \pi).$$

- Parzen kernel

$$k(z) = \begin{cases} 1 - 6z^2 + 6|z|^3 & \text{if } |z| \leq \frac{1}{2} \\ 2(1-|z|)^3 & \text{if } \frac{1}{2} < |z| \leq 1. \\ 0 & \text{otherwise.} \end{cases}$$

Its Fourier transform

$$K(u) = \frac{3}{8\pi}\left[\frac{\sin(u/4)}{u/4}\right]^4, \quad -\infty < u < \infty.$$

- Quadratic-Spectral kernel (i.e., Priestley)

$$k(z) = \frac{3}{(\pi z)^2}\left[\frac{\sin \pi z}{\pi z} - \cos(\pi z)\right], \quad -\infty < z < \infty.$$

Its Fourier transform

$$K(u) = \frac{3}{4\pi}[1 - (u/\pi)^2]\mathbf{1}(|u| \leq \pi).$$

5.3 Consistency of Kernel Spectral Estimator

Question: Why is the kernel spectral estimator $\hat{h}(\omega)$ consistent for $h(\omega)$?

We consider the integrated MSE criterion

$$\begin{aligned} IMSE(\hat{h}) &= E\int_{-\pi}^{\pi}\left|\hat{h}(\omega) - h(\omega)\right|^2 d\omega \\ &= E\int_{-\pi}^{\pi}\left|\hat{h}(\omega) - E\hat{h}(\omega)\right|^2 d\omega \\ &\quad + \int_{-\pi}^{\pi}\left|E\hat{h}(\omega) - h(\omega)\right|^2 d\omega. \end{aligned}$$

We first consider the bias of $\hat{h}(\omega)$.

Given $E[\hat{\gamma}(j)] = (1 - |j|/T)\gamma(j)$, we have

$$\begin{aligned}
E\hat{h}(\omega) - h(\omega) &= \frac{1}{2\pi}\sum_{j=1-T}^{T-1} k(j/p)E\hat{\gamma}(j)e^{-ij\omega} \\
&\quad -\frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma(j)e^{-ij\omega} \\
&= \frac{1}{2\pi}\sum_{j=1-T}^{T-1}[(1-|j|/T)k(j/p)-1]\gamma(j)e^{-ij\omega} \\
&\quad -\frac{1}{2\pi}\sum_{|j|>T-1}\gamma(j)e^{-ij\omega} \\
&= \frac{1}{2\pi}\sum_{j=1-T}^{T-1}[k(j/p)-1]\gamma(j)e^{-ij\omega} \\
&\quad -\frac{1}{2\pi T}\sum_{j=1-T}^{T-1}k(j/p)|j|\gamma(j)e^{-ij\omega} \\
&\quad -\frac{1}{2\pi}\sum_{|j|>T-1}\gamma(j)e^{-ij\omega} \\
&= -p^{-q}k_q h^{(q)}(\omega) + o(p^{-q}),
\end{aligned}$$

where $o(p^{-q})$ is uniform in $\omega \in [-\pi, \pi]$. Here, for the first term,

$$\begin{aligned}
&\frac{1}{2\pi}\sum_{j=1-T}^{T-1}[k(j/p)-1]\gamma(j)e^{-ij\omega} \\
&= -p^{-q}\frac{1}{2\pi}\sum_{j=1-T}^{T-1}\left\{\frac{[1-k(j/p)]}{|j/p|^q}\right\}|j|^q\gamma(j)e^{-ij\omega} \\
&= -p^{-q}k_q h^{(q)}(\omega)[1+o(1)]
\end{aligned}$$

as $p \to \infty$, where $k_q = \lim_{z\to 0}[1-k(z)]/|z|^q$, and the function

$$h^{(q)}(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}|j|^q\gamma(j)e^{-ij\omega}, \qquad \omega \in [-\pi, \pi],$$

is called the q-th order generalized derivative of $h(\omega)$. Note that $h^{(q)}(\omega)$ differs from the usual derivative. When q is even, we have

$$h^{(q)}(\omega) = \frac{1}{q!}\frac{d^q}{d\omega^q}h(\omega).$$

Note that a spectral peak will arise when $\gamma(j)$ decays to zero slowly as the lag order $j \to \infty$.

Next, we consider the second term of the bias. For the second term, we have

$$\begin{aligned}\frac{1}{2\pi T}\left|\sum_{j=1-T}^{T-1} k(j/p)\,|j|\,\gamma(j)e^{-ij\omega}\right| &\leq \frac{1}{2\pi T}\sum_{j=1-T}^{T-1}|j\gamma(j)| \\ &= O(T^{-1})\end{aligned}$$

if $\sum_{j=-\infty}^{\infty}|j\gamma(j)|<\infty$.

Similarly, for the last term of the bias,

$$\begin{aligned}\left|\sum_{|j|>T}\gamma(j)e^{-ij\omega}\right| &\leq \sum_{|j|>T}|\gamma(j)| \\ &\leq T^{-1}\sum_{|j|>T}|j\gamma(j)| \\ &= o(T^{-1})\end{aligned}$$

given $\sum_{j=-\infty}^{\infty}|j\gamma(j)|<\infty$, which implies $\sum_{|j|>T}|j\gamma(j)|\to 0$ as $T\to\infty$.

Thus, suppose $p^q/T\to 0$ such that $T^{-1}=o(p^{-q})$, which can be satisfied by choosing a suitable bandwidth $p=p(T)\to\infty$, we have

$$E\hat{h}(\omega)-h(\omega)=-p^{-q}k_q h^{(q)}(\omega)+o(p^{-q})$$

and the squared bias

$$\begin{aligned}&\int_{-\pi}^{\pi}\left[E\hat{h}(\omega)-h(\omega)\right]^2 d\omega \\ =\;& p^{-2q}k_q^2\int_{-\pi}^{\pi}\left[h^{(q)}(\omega)\right]^2 d\omega+o(p^{-2q}) \\ =\;& p^{-2q}k_q^2\frac{1}{2\pi}\sum_{j=-\infty}^{\infty}|j|^{2q}\gamma^2(j)+o(p^{-2q}).\end{aligned}$$

If $h^{(q)}(\omega)>0$, which can arise when (e.g.) the autocovariance $\gamma(j)$ is always positive for all j, then the bias is always negative. In other words, the kernel method always underestimates a spectral peak.

Unlike nonparametric estimation in time domain or space domain, there is no boundary bias problem, because $\hat{h}(\cdot)$ is a symmetric periodic function.

Next, for the integrated variance of $\hat{h}(\omega)$, we have

$$\begin{aligned} E[\hat{\gamma}(j) - E\hat{\gamma}(j)]^2 &= \text{var}[\hat{\gamma}(j)] \\ &= \frac{1}{T}\left[\sum_{j=-\infty}^{\infty} \gamma^2(j)\right][1 + o(1)] \\ &= \frac{1}{T}\int_{-\pi}^{\pi} h^2(\omega)d\omega\,[1 + o(1)]\,. \end{aligned}$$

Here, we have used the identity (see Priestley, 1981, p.xxx) that for any given lag order $j > 0$,

$$\text{var}[\hat{\gamma}(j)] = T^{-1}\sum_{m=1-(T-j)}^{T-j-1}\left(1 - \frac{|m|+j}{T}\right)\left[\gamma^2(m) + \gamma(m+j)\gamma(m-j) + \kappa_4(m,j,m+j)\right],$$

where $\kappa_4(i,j,k)$ is called the fourth order cumulant of the process $\{X_t\}$, defined as

$$\kappa_4(i,j,k) = E(X_tX_{t+i}X_{t+j}X_{t+k}) - E(\tilde{X}_t\tilde{X}_{t+i}\tilde{X}_{t+j}\tilde{X}_{t+k})$$

where $\{\tilde{X}_t\}$ is a Gaussian process with the same mean and autocovariance function as $\{X_t\}$. It follows that

$$\begin{aligned} & \frac{1}{(2\pi)}\sum_{j=1-T}^{T-1} k^2(j/p)\text{var}[\hat{\gamma}(j)] \\ = & \frac{1}{(2\pi)}\sum_{j=1-T}^{T-1} k^2(j/p)T^{-1}\sum_{m=1-(T-j)}^{T-j-1}\left(1 - \frac{|m|+j}{T}\right)\gamma^2(m) \\ & + \frac{1}{(2\pi)}\sum_{j=1-T}^{T-1} k^2(j/p)T^{-1}\sum_{m=1-(T-j)}^{T-j-1}\left(1 - \frac{|m|+j}{T}\right)\gamma(m+j)\gamma(m-j) \\ & + \frac{1}{(2\pi)}\sum_{j=1-T}^{T-1} k^2(j/p)T^{-1}\sum_{m=1-(T-j)}^{T-j-1}\left(1 - \frac{|m|+j}{T}\right)\kappa_4(m,j,m+j) \\ = & \hat{V}_1 + \hat{V}_2 + \hat{V}_3, \text{say,} \end{aligned}$$

where

$$\begin{aligned} \hat{V}_1 &= \frac{p}{T}\frac{1}{2\pi}\left[\sum_{m=-\infty}^{\infty}\gamma^2(m)\right]\left[\frac{1}{p}\sum_{j=1-T}^{T-1}k^2(j/p)\right][1 + o(1)] \\ &= \frac{p}{T}\int_{-\pi}^{\pi} h^2(\omega)d\omega\int_{-\infty}^{\infty}k^2(z)dz\,[1 + o(1)]\,, \end{aligned}$$

$$|\hat{V}_2| \le \frac{1}{T}\sum_{j=-\infty}^{\infty}|\gamma(j)|\sum_{m=-\infty}^{\infty}|\gamma(m)| = O(T^{-1})$$

if $\sum_{j=-\infty}^{\infty} |\gamma(j)| < \infty$, and finally, for the last term,

$$|\hat{V}_3| \leq \frac{1}{T} \sum_{j=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} |\kappa_4(i,j,k)| = O(T^{-1}).$$

if $\sum_{j=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} |\kappa_4(i,j,k)| < \infty$.

It follows that the IMSE

$$\begin{aligned}
\text{IMSE}(\hat{h}) &= \frac{p}{T} \int_{-\pi}^{\pi} h^2(\omega) d\omega \int_{-\infty}^{\infty} k^2(z) dz \\
&\quad + p^{-2q} k_q^2 \int_{-\pi}^{\pi} \left[h^{(q)}(\omega)\right]^2 d\omega \\
&\quad + o(p/T + p^{-2q}) \\
&= \frac{p}{T} \frac{1}{2\pi} \left[\sum_{j=-\infty}^{\infty} \gamma^2(j)\right] \int_{-\infty}^{\infty} k^2(z) dz \\
&\quad + \frac{1}{p^{2q}} k_q^2 \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} |j|^{2q} \gamma^2(j) \\
&\quad + o(T^{-1}p + p^{-2q}) \\
&= O(p/T + p^{-2q}).
\end{aligned}$$

Therefore, $\hat{h}(\omega)$ is consistent for $h(\omega)$ for any $\omega \in [-\pi, \pi]$ if $p \to \infty, p/T \to 0$ as $T \to \infty$. Differentiating the asymptotic IMSE$(\hat{h})$, we obtain the optimal bandwidth

$$\begin{aligned}
p_0 &= \left[\frac{2qk_q^2}{\int_{-\infty}^{\infty} k^2(z) dz} \frac{\int_{-\pi}^{\pi} [h^{(q)}(\omega)]^2 d\omega}{\int_{-\pi}^{\pi} h^2(\omega) d\omega}\right]^{\frac{1}{2q+1}} T^{\frac{1}{2q+1}} \\
&= c_0 T^{\frac{1}{2q+1}}, \text{ say.}
\end{aligned}$$

With this rate for p, the optimal convergence rate for $h(\omega)$ is $IMSE(\hat{h}) \propto T^{-\frac{2q}{2q+1}}$.

This optimal bandwidth is unknown because the optimal tuning parameter c_0 involves the unknown spectral density $h(\omega)$ and its generalized q-order derivative $h^{(q)}(\omega)$. Again, a plug-in or cross-validation method can be used.

It is shown that the optimal kernel is the Quadratic-Spectral kernel

$$k(z) = \frac{z}{(\pi z)^2} \left[\frac{\sin(\pi z)}{\pi z} - \cos(\pi z)\right], \quad -\infty < z < \infty.$$

Note that the Fourier transform of the QS kernel is the Epanechnikov kernel

$$K(u) = \frac{1}{4\pi} \left[1 - \left(\frac{u}{\pi}\right)^2\right] \mathbf{1}(|u| \leq \pi).$$

The latter is the optimal kernel for kernel density and regression estimation.

Question: Is there any equivalent expression for $\hat{h}(\omega)$ using the Fourier transform $K(u)$?

Yes, recall the formula

$$\hat{h}(\omega) = \frac{1}{2\pi} \sum_{t=1-T}^{T-1} k(j/p)\hat{\gamma}(j)e^{-ij\omega}, \qquad \omega \in [-\pi, \pi],$$

and the well-known result that the Fourier transform of the product between $\hat{\gamma}(j)$ and $k(j/p)$ is the convolution of their Fourier transforms, we can obtain

$$\begin{aligned}
\hat{h}(\omega) &= \frac{1}{2\pi} \sum_{j=1-T}^{T-1} k(j/p)\hat{\gamma}(j)e^{-\mathbf{i}j\omega} \\
&= \int_{-\pi}^{\pi} \hat{I}(\lambda) W_T(\omega - \lambda) d\lambda \\
&= \int_{-\pi}^{\pi} \hat{I}(\lambda) pK[p(\omega - \lambda)] d\lambda \\
&= \int_{-\pi}^{\pi} \hat{I}(\lambda) \frac{1}{h} K\left(\frac{\omega - \lambda}{h}\right) d\lambda,
\end{aligned}$$

where $h = p^{-1}$ is a bandwdith, $\hat{I}(\lambda)$ is the sample periodogram, namely,

$$\begin{aligned}
\hat{I}(\lambda) &= \frac{1}{2\pi T} \left| \sum_{t=1}^{T} X_t e^{\mathbf{i}t\lambda} \right|^2 \\
&= \frac{1}{2\pi} \sum_{j=1-T}^{T-1} \left(1 - \frac{|j|}{T}\right) \hat{\gamma}(j) e^{-\mathbf{i}j\lambda},
\end{aligned}$$

which is the discrete Fourier transform of $\hat{\gamma}(j)$, and the weighting function $W_T(\lambda)$ is the discrete Fourier transform of $k(j/p)$, namely,

$$\begin{aligned}
W_T(\lambda) &= \frac{1}{2\pi} \sum_{j=-(T-1)}^{T-1} k(j/p) e^{-\mathbf{i}j\lambda} \\
&= p \left[\frac{1}{2\pi p} \sum_{j=-\infty}^{\infty} k(j/p) e^{-\mathbf{i}(j/p)p\lambda} \right] \\
&= p \sum_{j=-\infty}^{\infty} K\left[p(\lambda + 2\pi j)\right] \\
&\sim pK(p\lambda).
\end{aligned}$$

Question: When

$$p\sum_{j=-\infty}^{\infty} K[p(\lambda+2\pi j)] = pK(p\lambda),\ \lambda\in[-\pi,\pi]?$$

When $K(\cdot)$ has bounded support on $[-\pi,\pi]$ and p is large, then the terms with $j\neq 0$ will all vanish to zero.

Since the periodogram $\hat{I}(\lambda)$ is the discrete Fourier transform of $\hat{\gamma}(j)$ and the weighting function $W_T(\lambda)$ is the discrete Fourier transform of $k(j/p)$, the Fourier transform of the product between $\hat{\gamma}(j)$ and $k(j/p)$ is the convolution of their Fourier transforms.

The weighting function $W_T(\lambda)$ plays a crucial role of local weighting and smoothing. We now provide a geometric interpretation of $\hat{h}(\omega)$.Put $h = p^{-1}$ so that $h\to 0$ as $T\to\infty$. Then

$$\begin{aligned}\hat{h}(\omega) &= \int_{-\pi}^{\pi}\hat{I}(\lambda)pK[p(\omega-\lambda)]d\lambda\\ &= \int_{-\pi}^{\pi}\hat{I}(\lambda)\frac{1}{h}K\left(\frac{\omega-\lambda}{h}\right)d\lambda.\end{aligned}$$

Therefore, $\hat{h}(\omega)$ is a smoothed version of the sample periodogram $\hat{I}(\lambda)$ over frequencies in a small neighborhood $[\omega-\pi h,\omega+\pi h]$, if $K(\cdot)$ has a support $[-\pi,\pi]$.

Suppose $q=2$. Then

$$E\hat{h}(\omega)-h(\omega) = -p^{-q}k_q h^{(q)}(\omega)+o(p^{-q}).$$

This can be relatively large when there is a spectral speak at frequency ω. Granger (1966) points out that the typical spectral shape of most economic time series is that it has a peak at frequency zero and then decays to zero as frequency increases.

Question: How to reduce the bias of $\hat{h}(\omega)$?

There are several approaches to reducing the bias of spectral density estimation. One popular approach is the pre-whitening procedure. Tukey (1957) and Andrews and Monahan (1992) consider this approach. To describe the basic idea, we consider an AR

approximation:

$$\begin{aligned} X_t &= \sum_{j=1}^{m} \psi_j X_{t-j} + u_t \\ &= \Psi(L) X_t + u_t. \end{aligned}$$

Then the innovation or residual $\{u_t\}$ will have weaker serial dependence. Put

$$\Psi(L) = 1 - \sum_{j=1}^{m} \psi_j L^j.$$

Then

$$u_t = \Psi(L) X_t$$

and it follows from Chapter 3 that the spectral density function of u_t

$$h_u(\omega) = |\Psi(e^{-\mathbf{i}\omega})|^2 h_X(\omega).$$

Thus, the spectral density of X_t is given by

$$h_X(\omega) = |\Psi(e^{-\mathbf{i}\omega})|^{-2} h_u(\omega).$$

In practice, we can first run a prewhitening regression, and obtain the parameter estimators $\{\hat{\psi}_j\}_{j=1}^m$. Then use the kernel method to estimate $h_u(\omega)$ using the prewhitening residual $\{\hat{u}_t\}$. Finally, obtain $\hat{h}_X(\omega) = |\hat{\Psi}(e^{-\mathbf{i}\omega})|^{-2} \hat{h}_u(\omega)$. This is called "recoloring". We note that the spectral density $h_u(\omega)$ of u_t is easier to estimate because it is "flatter" than the spectral density $h_X(\omega)$ of X_t.

For the prewhitening procedure, the bias can be reduced substantially but the variance is increased at the same time. As a result, MSE of $\hat{h}_X(\omega)$ may be larger than that without using prewhitening.

Another approach is to use the logarithmic transformation. Put

$$\lambda_k = \frac{2\pi k}{T} \text{ for } k = 0, ..., \left[\frac{T-1}{2}\right].$$

These are called Fourier frequencies. Then the sample periodogram of $\{X_t\}_{t=1}^T$

$$\hat{I}_X(\lambda_k) = f(\lambda_k) \hat{I}_\varepsilon(\hat{\lambda}) + \hat{R}_k,$$

where

$$\hat{I}_\varepsilon(\lambda_k) = \frac{1}{2\pi T} \left| \sum_{t=1}^{T} \varepsilon_t e^{it\lambda_k} \right|^2$$

is the sample periodogram of an innovation sequence $\{\varepsilon_t\}_{t=1}^T$, and R_k is an asymptotically negligible term. For $0 < k < [\frac{T-1}{2}]$.

A third approach is to use the wavelet approach, which can estimate the spectral peak more efficiently than the kernel method. For more discussions, see Härdle, Kerkyacharian, Picard and Tsybakov (1998), *Wavelets, Approximation and Statistical Applications,* Lecture Notes in Statistics Volume 129, Hong and Kao (2004, *Econometrica*), Hong and Lee (2001, *Econometric Theory*), and Lee and Hong (2001, *Econometric Theory*).

Kernel-based estimation for the spectral density function has been widely used in time series econometrics. The most well-known application is consistent estimation of the long-run variance-covariance matrix of a time series process (see Chapter 3). Observing that the long-run variance-covariance matrix is equal to 2π times the spectral density function at frequency zero, one can consistently estimate the long-run variance-covariance matrix by estimating the spectral density at frequency zero. Newey and West (1987, 1994) propose to use the Bartlett kernel to estimate the spectral density. Andrews (1991) propose a general class of kernel estimators for the long-run variance-covariance matrix, and show that the Quadratic-Spectral kernel is the optimal kernel that minimizes the asymptotic mean squared error of the kernel estimator. A key issue for kernel-based estimation of a long-run variance-covariance matrix is the choice of the smoothing parameter p. Newey and West (1994) and Andrews (1991) propose some data-driven plug-in methods to select p. In practice, kernel estimators often tend to display overrejections when used to construct test statistics. This arises particularly when data displays strong or persistent serial dependence. The main reason is that the kernel-based estimator tends to underestimate the true long-run variance-covariance matrix, as the asymptotic bias formula has indicated. Alternative approaches have been proposed, including the high-power kernel estimator (Phillips and Sun 1999), the so-called fixed b asymptotic method (Kiefer and Vogelsang 2005), and various self-normalization methods (e.g., Shao 2010, 2015). The basic idea of the self-normalization methods is to use a normalization factor that avoids choosing a smoothing parameter like p. The normalization factors are not consistent for the long-run variance-covariance matrix but are proportional to the long-run variance-covariance matrix up to a stochastic factor, which lead to a nonstandard asymptotic distribution for the proposed test statistics that is heavier than the normal distribution. These self-normalization methods can improve size of the proposed tests in finite sample, but they suffer from some power loss up to

various degrees, due to the heavy-tailed asymptotic distribution.

Another application of the kernel estimation of the spectral density function is to test for serial correlation of unknown form. In a time series linear regression model, it is often of interest to test the null hypothesis that the regression disturbance has no serial correlation of unknown form. Under the null hypothesis, the regression disturbance sequence is a white noise, and its spectral density function is a flat spectrum. Under the alternative hypothesis that there exists serial correlation, the spectral density of the disturbance is not flat and can be consistently estimated by the kernel method. Therefore, one can construct a consistent test for serial correlation of unknown form by comparing a kernel-based spectral density estimator with the flat spectrum. To compare the two spectral density estimators under the null and alternative hypotheses, one can use the squared L_2-norm, the squared Hellinger metric, or the Kullback-Leibler information criterion. As the sample size $T \to \infty$, these distance or divergence measures vanish to zero under the null hypothesis and converge to nonzero limits under the alternative hypothesis, giving the tests their asymptotic unit power. Therefore, if these distance or divergence measures are close to zero, then the null hypothesis holds; if they are not close to zero, then the alternative hypothesis must be true. How large these distance or divergence measures should be in order to be considered as significantly large is determined by their sampling distributions. For more discussion, see Hong (1996) or Chapter 7 for the L_2-norm-based test.

6 Conclusion

In this chapter, we have introduced some smoothed nonparametric estimation methods in both time domain and frequency domain. The functions of interest include but are not restricted to probability density functions, regression functions, trend functions of time, time-varying functional coefficients, and spectral density functions. Associated with the first three functions of interest are nonparametric methods in time domain, and associated with the spectral density estimation are nonparametric methods in frequency domain.

Nonparametric methods can be divided into two categories: global smoothing and local smoothing. In this chapter, we consider local smoothing. In particular, we introduce the kernel smoothing methods for density function, regression function and spectral density functions, and local polynomial smoothing methods for regression func-

tion, functional coefficient models, functional coefficient models in a linear or nonlinear setup. Relationships between these methods are also discussed, including the Fourier transform relationship between the nonparametric estimation methods in time domain and frequency domain. Nonparametric methods have been widely applied in economics and finance.

EXERCISE 6

6.1. What are the main advantages of nonparametric smoothing methods in time series econometrics? Why have nonparametric methods become popular in recent years?

6.2. What is the boundary problem for the kernel smoothing method? How can one alleviate this boundary problem?

6.3. What is the curse of dimensionality associated with nonparametric smoothing?

6.4. Why can the local linear smoother automatically solve for the boundary bias problem in nonparametric regression estimation?

6.5. Suppose $\{X_t\}_{t=1}^T$ is an IID random sample with a twice continuously differentiable marginal density function $g(x)$ on support $[a, b]$. Define the kernel density estimator

$$\hat{g}(x) = \frac{1}{T}\sum_{t=1}^{T} K_h(x - X_t),$$

where $K_h(x - X_t) = h^{-1}K[(x - X_t)/h]$, $K(\cdot)$ is a positive kernel with support on $[-1, 1]$, and $h = h(T) \to 0$ is a bandwidth.

(1) For $x \in [a + h, b - h]$, derive the asymptotic bias expression for $E\hat{g}(x) - g(x)$.

(2) For $x \in [a + h, b - h]$, derive the asymptotic variance expression for $\text{var}(\hat{g}(x)) = E[\hat{g}(x) - E\hat{g}(x)]^2$.

(3) Find the asymptotic expression for the mean squared error MSE $E[\hat{g}(x) - g(x)]^2$.

(4) Derive the optimal bandwidth h^* that maximizes the asymptotic MSE of $\hat{g}(x)$.

(5) What is the asymptotic MSE when evaluated at the optimal bandwidth h^*.

6.6. Suppose $K(\cdot)$ is a higher order (q-th order) kernel such that $\int_{-1}^{1} K(u)du = 1, \int_{-1}^{1} u^j K(u)du = 0$ for $1 \le j \le q - 1, \int_{-1}^{1} u^q K(u)du = C_K(q)$ and $\int_{-1}^{1} K^2(u)du = D_K$. In addition, assume that $g(x)$ is q-time continuously differentiable on $[a, b]$. Answer (1)–(5) in Exercise 6.5 again.

6.7. In the setup of Exercise 6.5, further assume $g(x) \ge \epsilon > 0$ for some constant $\epsilon > 0$. Consider the asymptotic bias of $\hat{g}(x)$ for $x = a + \rho h \in [a, a + h]$ for $\rho \in [0, 1)$.

(1) Show that $\sup_{x\in[a,a+h]} |E\hat{g}(x) - g(x)|$ never vanishes to zero as $h \to \infty$.

(2) There are several approaches to deal with the boundary bias problem in (1). One simple way is to consider the following kernel estimator

$$\hat{g}(x) = \frac{1}{T}\sum_{t=1}^{T} K_h(x, X_t),$$

where

$$K_h(x,y) \equiv \begin{cases} h^{-1}K\left(\frac{x-y}{h}\right) / \int_{-(x/h)}^{1} K(u)du, & \text{if } \ x \in [0,h), \\ h^{-1}K\left(\frac{x-y}{h}\right), & \text{if } \ x \in [h,1-h], \\ h^{-1}K\left(\frac{x-y}{h}\right) / \int_{-1}^{(1-x)/h} K(u)du, & \text{if } \ x \in (1-h,1] \end{cases}$$

and $K(\cdot)$ is a standard kernel. This estimator differs from the estimator in Exercise 6.5 in the boundary regions but not in the interior regions. Show that $\sup_{x\in[a,a+h)} |E\hat{g}(x) - g(x)| \to 0$ as $h \to 0$.

6.8. One method to deal with the boundary bias problem of kernel estimation is to use the so-called reflection method. This method constructs the kernel density estimate based on the "reflected" data $\{-X_t\}_{t=1}^T$ and the original data $\{X_t\}_{t=1}^T$. Suppose X_t has a twice-continuously differentiable marginal pdf $g(x)$ with the support $[a,b]$, and x is a left boundary point in $[a,a+h)$ and $x \geq 0$. Then the reflection method uses an estimator

$$\hat{g}(x) = \frac{1}{T}\sum_{t=1}^{T} K_h(x - X_t) + \frac{1}{T}\sum_{t=1}^{T} K_h[x - (-(X_t - a))],$$

where $K_h(x - X_t) = h^{-1}K[(x - X_t)/h]$, $K : [-1,1] \to R^+$ is a pre-specified symmetric pdf with support $[-1,1]$ and h is the bandwidth. Find the bias $E\hat{g}(x) - g(x)$ for the following cases:

(1) $x \in [a, a+h)$;

(2) $x \in [ah, b-h]$.

6.9. Suppose a second order kernel $K : [-1,1] \to R$ is symmetric about zero, $\int_{-1}^{1} u^2K(u) = C_K < \infty$, $\int_{-1}^{1} K^2(u)du = D_K < \infty$, but $\int_{-1}^{1} K(u)du \neq 1$.

(1) Derive the bias of the Nadaraya-Watson estimator for $r(x)$, where x is a given point in the interior region.

(2) Derive the bias of the Nadaraya-Watson estimator when x is in the boundary region.

For both (1) and (2), explain the results you obtained and compare them to the results obtained under the condition that $\int_{-1}^{1} K(u)du = 1$.

6.10. Suppose a data generating process is given by

$$Y_t = 1 + X_t - 0.25X_t^2 + \varepsilon_t, \qquad t = 1, ..., T,$$

where $\{X_t\} \sim$ IID $U[0, 2\sqrt{3}]$, $\{\varepsilon_t\} \sim$ IID $N(0,1)$, and $\{X_t\}$ and $\{\varepsilon_t\}$ are mutually independent.

(1) Generate a data $\{Y_t, X_t\}_{t=1}^T$ with $T = 200$ using a random number generator on a computer, and plot the sample point on the xy-plane, and plot the true regression function $r(x) = E(Y_t|X_t = x)$.

(2) Use a Nadaraya-Watson estimator to estimate the regression function $r(X_t) = E(Y_t|X_t)$ on 100 equally spaced grid points on $[0, 2\sqrt{3}]$. Use the quatic kernel $K(u) = \frac{15}{16}(1 - |u|^2)^2 1(|u| \leq 1)$ and choose the bandwidth $h = S_X T^{-\frac{1}{5}}$, where S_X is the sample standard deviation of $\{X_t\}_{t=1}^T$. Plot the estimator $\hat{r}(x)$ on the xy-plane.

(3) Use a local linear estimator to estimate the regression function $r(x)$ on 100 equally spaced grid points on $[0, 2\sqrt{3}]$, with the same kernel $K(\cdot)$ and bandwidth h as in part (2). Plot the estimator for $r(x)$ on the xy-plane.

6.11. Again, in the setup of Exercise 6.5, further assume $g(x) \geq \epsilon > 0$ for some constant $\epsilon > 0$. Consider the asymptotic bias of $\hat{g}(x)$ for $x = a + \rho h \in [a, a + h]$ for $\rho \in [0, 1)$. Another method to deal with the boundary bias problem is to use the so-called jackknife kernel method.

(1) For $x = a + \rho h \in [a, a + h)$, we consider an estimator

$$\bar{g}(x) = \hat{g}(x; h) + \beta \left[\hat{g}(x; h) - \hat{g}(x; \alpha h)\right],$$

where

$$\begin{aligned} \hat{g}(x; h) &= \frac{1}{T} \sum_{t=1}^{T} h^{-1} K_\rho \left(\frac{x - X_t}{h} \right), \\ K_\rho(u) &\equiv \frac{K(u)}{\omega_K(0, \rho)}, \end{aligned}$$

and $\omega_K(i, \rho) = \int_{-\rho}^{1} u^i K(u) du$ for $i = 0, 1, 2$.

Now define a new kernel (called jackknife kernel)

$$K_\rho^J(u) = (1 + \beta) K_\rho(u) - \frac{\beta}{\alpha} K_{\frac{\rho}{\alpha}}(\frac{u}{\alpha})$$

where β is the same as in $\bar{g}(x)$. Show that

$$\bar{g}(x) = \frac{1}{T} \sum_{t=1}^{T} h^{-1} K_\rho^J \left(\frac{x - X_t}{h} \right).$$

(2) Find the expression for β in terms of $\omega_K(\cdot, \rho)$ and α such that $\sup_{x \in [a, a+h)} |E\bar{g}(x) - g(x)| = O(h^2)$.

(3) Suppose now $x = b - \rho h \in (b-h, b]$. Can we use $\bar{g}(x)$ and get an asymptotic bias of order $O(h^2)$. If yes, verify it; if not, derive an estimator so that you can obtain an $O(h^2)$ bias for $x \in (b-h, b]$.

6.12. Suppose $\{X_t\}_{t=1}^T$ is an IID random sample with a twice continuously differentiable marginal density function $g(x)$ on support $[a, b]$. Define the kernel density estimator

$$\hat{g}(x) = \frac{1}{T}\sum_{t=1}^{T} K_h(x - X_t),$$

where $K_h(x - X_t) = h^{-1}K[(x - X_t)/h]$, $K(\cdot)$ is a standard kernel (usually called second order kernel or positive kernel) with support on $[-1, 1]$ such that , and $h = h(T) \to 0$ is a bandwidth.

(1) For $x \in [a + h, b - h]$, derive the asymptotic bias expression for $E[\hat{g}(x)] - g(x)$.

(2) For $x \in [a + h, b - h]$, derive the asymptotic variance expression for $\text{var}[\hat{g}(x)] = E\left\{\hat{g}(x) - E[\hat{g}(x)]\right\}^2$.

(3) Find the asymptotic expression for the mean squared error MSE $E[\hat{g}(x) - g(x)]^2$.

(4) Derive the optimal bandwidth h^* that maximizes the asymptotic MSE of $\hat{g}(x)$.

(5) What is the asymptotic MSE when evaluated at the optimal bandwidth h^*.

6.13. Suppose $K(\cdot)$ is a higher order (q-th order) kernel such that $\int_{-1}^{1} K(u)du = 1, \int_{-1}^{1} u^j K(u)du = 0$ for $1 \le j \le q-1, \int_{-1}^{1} u^q K(u)du = C_K(q)$ and $\int_{-1}^{1} K^2(u)du = D_K$. In addition, assume that $g(x)$ is q-time continuously differentiable on $[a, b]$. Answer (1)–(5) in Exercise 6.12 again.

6.14. In the setup of Exercise 6.12, further assume $g(x) \ge \epsilon > 0$ for some constant $\epsilon > 0$. Consider the asymptotic bias of $\hat{g}(x)$ for $x = a + \rho h \in [a, a + h]$ for $\rho \in [0, 1)$.

(1) Show that $\sup_{x \in [a, a+h]} |E\hat{g}(x) - g(x)|$ never vanishes to zero as $h \to \infty$.

(2) There are several approaches to deal with the boundary bias problem in (1). One simple way is to consider the following kernel estimator

$$\hat{g}(x) = \frac{1}{T}\sum_{t=1}^{T} K_h(x, X_t),$$

where

$$K_h(x, y) \equiv \begin{cases} h^{-1}K\left(\frac{x-y}{h}\right) / \int_{-(x/h)}^{1} K(u)du, & \text{if } \ x \in [0, h), \\ h^{-1}K\left(\frac{x-y}{h}\right), & \text{if } \ x \in [h, 1-h], \\ h^{-1}K\left(\frac{x-y}{h}\right) / \int_{-1}^{(1-x)/h} K(u)du, & \text{if } \ x \in (1-h, 1] \end{cases}$$

and $K(\cdot)$ is a standard kernel. This estimator differs from the estimator in Exercise 6.12 in the boundary regions but not in the interior regions. Show that $\sup_{x\in[a,a+h)} |E[\hat{g}(x)] - g(x)| \to 0$ as $h \to 0$.

6.15. One method to deal with the boundary bias problem of kernel estimation is the so-called reflection method. This method constructs the kernel density estimate based on the "reflected" data $\{-X_t\}_{t=1}^T$ and the original data $\{X_t\}_{t=1}^T$. Suppose X_t has a twice-continuously differentiable marginal pdf $g(x)$ with the support $[a, b]$, and x is a left boundary point in $[a, a+h)$ and $x \geq 0$. Then the reflection method uses an estimator

$$\hat{g}(x) = \frac{1}{T}\sum_{t=1}^{T} K_h(x - X_t) + \frac{1}{T}\sum_{t=1}^{T} K_h[x - (-(X_t - a))],$$

where $K_h(x - X_t) = h^{-1}K[(x - X_t)/h]$, $K : [-1, 1] \to R^+$ is a pre-specified symmetric pdf with support $[-1, 1]$ and h is the bandwidth. Find the bias $E\hat{g}(x) - g(x)$ for the following two cases:

(1) $x \in [a, a + h)$;

(2) $x \in [ah, b - h]$.

6.16. Suppose a data generating process is given by

$$Y_t = 1 + X_t - 0.25X_t^2 + \varepsilon_t, \qquad t = 1, ..., T,$$

where $\{X_t\} \sim$ IID $U[0, 2\sqrt{3}]$, $\{\varepsilon_t\} \sim$ IID $N(0, 1)$, and $\{X_t\}$ and $\{\varepsilon_t\}$ are mutually independent.

(1) Generate a data $\{Y_t, X_t\}_{t=1}^T$ with $T = 200$ using a random number generator on a computer, and plot the sample point on the xy-plane, and plot the true regression function $r(x) = E(Y_t|X_t = x)$.

(2) Use a Nadaraya-Watson estimator to estimate the regression function $r(X_t) = E(Y_t|X_t)$ on 100 equally spaced grid points on $[0, 2\sqrt{3}]$. Use the quartic kernel $K(u) = \frac{15}{16}(1 - |u|^2)^2 1(|u| \leq 1)$ and choose the bandwidth $h = S_X T^{-\frac{1}{5}}$, where S_X is the sample standard deviation of $\{X_t\}_{t=1}^T$. Plot the estimator $\hat{r}(x)$ on the xy-plane.

(3) Use a local linear estimator to estimate the regression function $r(x)$ on 100 equally spaced grid points on $[0, 2\sqrt{3}]$, with the same kernel $K(\cdot)$ and bandwidth h as in part (2). Plot the estimator for $r(x)$ on the xy-plane.

6.17. Again, in the setup of Exercise 6.12, further assume $g(x) \geq \epsilon > 0$ for some constant $\epsilon > 0$. Consider the asymptotic bias of $\hat{g}(x)$ for $x = a + \rho h \in [a, a + h]$ for

$\rho \in [0,1)$. Another method to deal with the boundary bias problem is to use the so-called jackknife kernel method.

(1) For $x = a + \rho h \in [a, a+h)$, we consider $\bar{g}(x) = \hat{g}(x;h) + \beta\left[\hat{g}(x;h) - \hat{g}(x;\alpha h)\right]$, where

$$\begin{aligned} \hat{g}(x;h) &= \frac{1}{T}\sum_{t=1}^{T} h^{-1} K_\rho\left(\frac{x - X_t}{h}\right), \\ K_\rho(u) &\equiv \frac{K(u)}{\omega_K(0,\rho)}, \end{aligned}$$

and $\omega_K(i,\rho) = \int_{-\rho}^{1} u^i K(u)du$ for $i = 0, 1, 2$.

Now, define a new kernel (called jackknife kernel)

$$K_\rho^J(u) = (1+\beta)K_\rho(u) - \frac{\beta}{\alpha}K_{\frac{\rho}{\alpha}}(\frac{u}{\alpha})$$

where β is the same as in $\bar{g}(x)$. Show that

$$\bar{g}(x) = \frac{1}{T}\sum_{t=1}^{T} h^{-1} K_\rho^J\left(\frac{x - X_t}{h}\right).$$

(2) Find the expression for β in terms of $\omega_K(\cdot,\rho)$ and α such that $\sup_{x\in[a,a+h)} |E[\bar{g}(x)] - g(x)| = O(h^2)$.

(3) Suppose now $x = b - \rho h \in (b-h, b]$. Can we use $\bar{g}(x)$ and get an asymptotic bias of order $O(h^2)$. If yes, verify it; if not, derive an estimator so that you can obtain an $O(h^2)$ bias for $x \in (b-h, b]$.

6.18. Derive the asymptotic MSE formula for the local polynomial time-trend estimator $\hat{m}(t_0/T)$ for t_0 in the boundary region $[1, Th)$ or $(T - Th, T]$.

6.19. Derive the asymptotic distribution for the local polynomial time-trend estimator $\hat{m}(t_0/T)$ for t_0 in the boundary region $[1, Th)$ or $(T - Th, T]$.

6.20. Derive the asymptotic MSE formula for the local polynomial estimator $\hat{\alpha}(t_0/T)$ in the locally linear regression model, where t_0 is in the interior region $[Th, T - Th]$.

6.21. Derive the asymptotic MSE formula for the local polynomial estimator $\hat{\alpha}(t_0/T)$ in a locally linear time-varying regression model, where t_0 is in the boundary region region $[1, Th)$ or $(T - Th, T]$.

6.22. Suppose $\{X_t\}$ is a sixth-order stationary time series process. Define a kernel-based bispectral density estimator

$$\hat{b}(\omega_1, \omega_2) = \frac{1}{(2\pi)^2} \sum_{j=1-T}^{T-1} \sum_{l=1-T}^{T-1} k(j/p)k(l/p)k[(j-l)/p]\hat{C}(0, j, l)e^{ij\omega_1 + il\omega_2}.$$

(1) If $k_2 = \lim \frac{1-k(z)}{|z|^2} \in (0, \infty)$, then show the bias

$$E\left[\hat{b}(\omega_1, \omega_2)\right] - b(\omega_1, \omega_2) = -\frac{1}{2}\frac{k_2}{p^2} D^{(2)}(\omega_1, \omega_2)[1 + o(1)],$$

where

$$D^2(\omega_1, \omega_2) = \left(\frac{\partial^2}{\partial\omega_1^2} - \frac{\partial^2}{\partial\omega_1\partial\omega_2} + \frac{\partial^2}{\partial\omega_2^2}\right) b(\omega_1, \omega_2).$$

See Subba Rao and Gabr (1984) for the derivation of the bias.

(2) Show

$$\text{var}\left[\hat{b}(\omega_1, \omega_2)\right] = \frac{p^2}{T}\frac{V}{2\pi} h(\omega_1)h(\omega_2)h(\omega_1 + \omega_2)\left[1 + o(1)\right],$$

where

$$V = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} k^2(u)k^2(v)k^2(u-v)dudv.$$

See Brillinger and Rosenblatt (1967a) for the derivation of the variance of the bispectral density estimator.

(3) Obtain the conditions on bandwidth $p = p(T)$ so that $\hat{b}(\omega_1, \omega_2)$ is consistent for $b(\omega_1, \omega_2)$ as the sample size $T \to \infty$.

6.23. Suppose $\{X_t\}$ is a strictly stationary time series. Define a generalized spectral density estimator

$$\hat{f}(\omega, u, v) = \frac{1}{2\pi} \sum_{j=1-T}^{T-1} (1 - |j|/T)^{1/2} k(j/p)\hat{\sigma}_j(u, v)e^{-ij\omega},$$

where

$$\hat{\sigma}_j(u, v) = \hat{\varphi}_j(u, v) - \hat{\varphi}_j(u, 0)\hat{\varphi}_j(0, v),$$

and

$$\hat{\varphi}_j(u, v) = (T - |j|)^{-1} \sum_{t=|j|+1}^{T} e^{iuX_t + ivX_{t-|j|}}$$

is the empirical characteristic function of $(X_t, X_{t-|j|})$.

(1) Suppose $0 < k_q < \infty$, where $k_q = [1 - k(z)]/|z|^q$. Derive the asymptotic bias of $\hat{f}(\omega, u, v)$.

(2) Derive the asymptotic variance of $\hat{f}(\omega, u, v)$.

(3) Derive the conditions on the smoothing parameter p so that $\hat{f}(\omega, u, v)$ is consistent for $f(\omega, u, v)$.

(4) Derive the optimal smoothing parameter p_0 that minimizes the integrated MSE of $\hat{f}(\omega, u, v)$.