Nonparametric Methods in Time Series

Yongmiao Hong

Cornell University

May 2012

Abstract: So far, we have considered a variety of measures for serial dependence and have focused on their probabilistic properties. How these measures can be consistently estimated has not been considered. We will introduce some popular nonparametric methods (particularly the kernel smoothing method) to estimate functions of interest, such as probability density functions, autoregression functions, power 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 kernel-based estimators in various contexts.

Key words: Smoothing, Taylor series expansion, density function, generalized spectral density, local polynomial smoothing, kernel method, regression function, spectral density
References:

Nonparametric Methods in Time Domain

Silverman, B. (1986): *Nonparametric Density Estimation and Data Analysis.* Chapman and Hall: London.
Hardle, 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.

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.

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

- 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 power spectrum $h(\omega)$ 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 nonparametric nonlinear autoregressive conditional heteroskedastic process

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

  where $\mu(\cdot)$ and $\sigma(\cdot)$ are unknown functions. 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 seminonparametric 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 $\mu_j(\cdot)$ are unknown?
- How to use these estimators in economic and financial applications?

Remark: Nonparametric smoothing first arose from spectral density estimation in time series. In a discussion of the seminal paper by Bartlett (1946), Henry E. Daniels suggested that a possible improvement on spectral density estimation could be made by smoothing the periodogram. 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 half a century ago.

Remark: In the earlier stage of nonlinear time series analysis, focus was on various nonlinear parametric forms. 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.

Question: Why is the nonparametric method popular in statistics and econometrics?
Answer:

(i) Demands for nonlinear approaches;

(ii) Availability of large data sets;
(iii) Advance in computer technology.

Granger (1999): The speed in computing technology increases much faster than the speed at which data grows.

1 Motivation

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

Motivating example 1 [regression function]

For simplicity, we consider $r_1(x) = E(X_t|X_{t-1} = x)$. We write

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

where $E(\varepsilon_t|X_{t-1}) = 0$ a.s. 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

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

where the Fourier coefficient

$$\alpha_j = \int r_1(x)\psi_j(x)dx,$$

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

Example 1: Suppose $r_1(x) = x^2$ where $x \in [-\pi, \pi]$. Then

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

Example 2: Suppose

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

Then

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

Because $r_1(x)$ is square-integrable, we have

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

where $\delta_{j,k} = 1$ if $j = k$ and 0 otherwise.

Remark: $\alpha_j \to 0$ as $j \to \infty$. That is, α_j becomes less important as the order j grows to infinity.

This suggests that a truncated sum

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

can 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 $p \to \infty$.

Difficulty: However, the coefficient α_j is unknown.

To obtain a feasible estimator for $r_1(x)$, we consider the following sequence of regression models

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

where $p \equiv p(T)$ is the number of series terms that depends on the sample size T. We need p/T goes to 0. Note that ε_{pt} is not the same as the true error ε_t for each given p. Instead, it contains the true error ε_t and the bias $b_p(X_t)$.

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

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

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.

Remark: $\{\psi_j(\cdot)\}$ can be the Fourier series, i.e., the sin and cosine functions. See (e.g.) Andrews (1991, *Econometrica*), Hong and White (1995, *Econometrica*).

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

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

where $\phi(x) = (2\pi)^{-1/2} \exp(-\frac{1}{2}x^2)$ is the $N(0,1)$ density, 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 $\sum_{j=0}^{\infty} \beta_j^2 < \infty$.

Remark: 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 densities

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

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. Suppose $\{X_t\}$ is an i.i.d. sample. Then

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

To ensure that $\hat{g}_p(x) = \hat{C}(p)\phi(x)\Sigma_{j=0}^{p}\hat{\beta}_j H_j(x)$ is asymptotically unbiased, we must let $p = p(T) \to \infty$ as $T \to \infty$. However, p must grow slowly so that the sampling variation of $\hat{\beta}$ will not be too large.

Remark: For the use of Hermite Polynomial series expansions, see (e.g.) Gallant and Tauchen (1996, *Econometric Theory*) and Ait-Sahalia (2002, *Econometrica*).

Question: What is the advantage of nonparametric estimation methods?

They require few assumptions or restrictions on the data generating process. In particularly, 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 it is linear or nonlinear. Thus, nonparametric methods can reduce potential systematic biases which are more likely to encounter in parametric models.

Question: What is the disadvantage of nonparametric methods?

(i) They require a large data set for reasonable estimation. There exists a problem of "curse of dimensionality", which will be explained below.

(ii) Coefficients are usually difficult to interpret from an economic point of view;

(iii) There exists a danger of potential overfitting.

Remark: The two motivating examples are the so-called orthogonal series expansion methods. There are many other nonparametric methods, such as kernel methods and local polynomials methods. Series expansion methods are called **global smoothing** methods, because the coefficients are estimated using all observations, and they are then used to evaluate the value of the underlying function at each point. In contrast, kernel and local polynomial methods are called **local smoothing** methods, because the estimation only requires the observations in a neighborhood of the point of interest. Below we will mainly focus on kernel smoothing methods, due to its simplicity and intuitive nature.

Remark: A nonparametric model is an increasing sequence of parametric models, as the sample size T grows.

2 Kernel Density Method

2.1 Kernel estimation of $g(x)$

Basic Question: How to estimate the marginal pdf $g(x)$ of X_t?

Parametric Approach: Suppose $\{X_t\}$ is a time series process with the same marginal density function $g(x)$. Assume that $g(x)$ is a $N(\mu, \sigma^2)$ pdf with unknown μ and σ^2. Then we know the functional form of $g(x)$ up to two unknown parameters $\theta = (\mu, \sigma^2)'$:

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

To estimate $g(x|\theta)$, it suffices to estimate μ and σ^2. We can use the maximum likelihood estimation (MLE) method:

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

Remark: $\sqrt{T}(\hat{\theta} - \theta_0) = O_P(1)$, or $\hat{\theta} - \theta_0 = O_P(T^{-1/2})$, where $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)'$ and $\theta_0 = (\mu_0, \sigma_0^2)'$ is the true parameter.

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

Definition [Convergence in probability]: Let $\{\delta_T, T \geq 1\}$ be a sequence of positive numbers. A random variable Y_T is said to be 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

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

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

$$\begin{aligned} g(x|\hat{\theta}) - g(x) &= g(x|\theta_0) - g(x) \\ &\quad + \frac{\partial}{\partial\theta} g(x|\bar{\theta})(\hat{\theta} - \theta_0) \\ &= \text{approximation error (bias)} \\ &\quad + \text{sampling error} \\ &= 0 + \frac{1}{\sqrt{T}} \frac{\partial}{\partial\theta} g(x|\bar{\theta}) \sqrt{T}(\hat{\theta} - \theta_0) \\ &= 0 + O_P(T^{-1/2}) \\ &= O_P(T^{-1/2}). \end{aligned}$$

Question: What happens if the Gaussianity assumption fail? That is, what happens if $g(x|\theta) \neq g(x)$ for all θ?

Answer: $g(x|\hat{\theta})$ will not be consistent for $g(x)$ because the bias $g(x|\theta) - g(x)$ never vanishes, where $\theta = p\lim \hat{\theta}$.

Nonparametric Kernel Approach:

Basic idea of local smoothing: The purpose of nonparametric probability density estimation is to construct an estimate of a probability density function without imposing structural assumptions. Typically the only conditions imposed on the probability density function are that it has at least two bounded derivatives. In this circumstance we may use only local information about the value of the density at any given point. That is, the value of the density of a point x must be calculated from data values that lie in the neighborhood of x, and to ensure consistency the neighborhood must shrink as the sample size increases. In the case of kernel-type density estimation, the radius of the effective neighborhood is roughly equal to the "bandwidth" or the smoothing parameter of the estimator. Under the assumption that the density is univariate with at least two bounded derivatives, and using a nonnegative kernel function, the size of bandwidth that optimizes the performance of the estimator in the mean squared error criterion is $T^{-1/5}$. The number of "parameters" needed to model the unknown density 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, this number growing with increasing n.

There are two basic instruments in kernel estimation: the kernel function $K(\cdot)$ and the bandwidth h. 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.

2.1.1 Kernel Function

Kernel Function $K(\cdot)$: A positive kernel function $K(\cdot)$ is a pre-specified symmetric pdf such that (i) $\int_{-\infty}^{\infty} K(u)du = 1$, (ii) $\int_{-\infty}^{\infty} K(u)udu = 0$, (iii) $\int_{-\infty}^{\infty} u^2K(u)du = C_k < \infty$, (iv) $\int_{-\infty}^{\infty} K^2(u)du = D_k$.

Remarks:

(i) $K(\cdot)$ is a weighting function that will "discount" the observations whose values are more away from the point x that we are interested in.

(ii) The kernels satisfying the above condition are called a second order kernel or positive kernel. More generally, we can define a q-th order kernel $K(\cdot)$, where $q \geq 2$: $K(\cdot)$ satisfies the conditions that $\int K(u)du = 1, \int u^jK(u)du = 0$ for $1 \leq j \leq q-1$, $\int u^qK(u)du < \infty$ and $\int K^2(u)du < \infty$. For a higher order kernel ($q > 2$), $K(\cdot)$ 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?)

Examples of second order kernel $K(\cdot)$:

- 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
$$K(u) = \frac{3}{4}(1-u^2)\mathbf{1}(|u| \leq 1).$$
- Quatic kernel
$$K(x) = \frac{15}{16}(1-u^2)^2\mathbf{1}(|u| \leq 1).$$

2.1.2 Consistency of the Kernel Density Estimator for $g(x)$

Question: How does the kernel method work?

Given a pre-chosen kernel $K(u)$, we can define a kernel density estimator for $g(x)$ using $\{X_t\}_{t=1}^T$:

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

where $K_h(u) = h^{-1}K(u/h)$, $h > 0$ is called a bandwidth or a window size, and $\hat{F}(y) = T^{-1}\Sigma_{t=1}^{T} 1(X_t \leq t)$ is the empirical distribution function.

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

$$\begin{aligned}\hat{g}(x) &= \frac{1}{2hT}\sum_{t=1}^{T}\mathbf{1}(|x - X_t| \leq h) \\ &= \text{the relative sample frequency of the observations} \\ &\quad \text{on the interval } [x-h, x+h].\end{aligned}$$

Remark: $2hT$ is approximately the sample size of the interval $[x-h, x+h]$.

Question: Under what conditions will $\hat{g}(x)$ be consistent for $g(x)$?

Assumption: (i) $\{X_t\}$ is a strictly stationary process with marginal *pdf* $g(x)$. (ii) $g(x)$ has a 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 $x_1, x_2 \in [a, b]$ and $C < \infty$.

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

Remark: By convention, the derivatives of $g(\cdot)$ at 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 $g''(a)$ and $g''(b)$.

For convenience, we further assume:

Assumption: $K(u)$ has a bounded support on $[-1, 1]$.

This is not necessary, but it simplifies the asymptotic analysis below. Now we decompose

$$\begin{aligned}\hat{g}(x) - g(x) &= [E\hat{g}(x) - g(x)] + [\hat{g}(x) - E\hat{g}(x)] \\ &= \text{bias} + \text{sampling error.}\end{aligned}$$

The first term is the bias, which is nonstochastic.

For any x in the interior region $[a+h, b-h]$ of the support $[a,b]$ of X_t, we have

$$\begin{aligned}
& E\hat{g}(x) - g(x) \\
= \; & \frac{1}{T}\sum_{t=1}^{T} EK_h(x - X_t) - g(x) \\
= \; & EK_h\left(x - X_t\right) - g(x) \text{ (by identical distribution)} \\
= \; & \int_a^b \frac{1}{h} K\left(\frac{x-y}{h}\right) g(y)dy - g(x) \\
= \; & \int_{(a-x)/h}^{(b-x)/h} K(u)g(x+hu)du - g(x) \text{ (by change of variable } \frac{y-x}{h} = u) \\
= \; & \int_{-1}^{1} K(u)g(x+hu)du - g(x) \\
= \; & g(x)\int_{-1}^{1} K(u)du - g(x) \\
& + hg'(x)\int_{-1}^{1} uK(u)du \\
& + \frac{1}{2}h^2 \int_{-1}^{1} u^2 K(u) g''(x+\lambda hu)du \\
& \text{(by a 2nd order Taylor expansion, where } \lambda \in (0,1)) \\
= \; & \frac{1}{2}h^2 C_K g''(x) + \frac{1}{2}h^2 \int_{-1}^{1} [g''(x+\lambda hu) - g''(x)]u^2 K(u)du \\
= \; & \frac{1}{2}h^2 C_K g''(x) + o(h^2)
\end{aligned}$$

where the second term

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

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

Remarks:

(i) For 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 $T \to \infty$.

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

(iii) **Question**: What happens if x is outside the interior region $[a+h, b-h]$?

Suppose $x = a + \lambda h \in [a, a+h)$, where $\lambda \in [0, 1)$. Then

$$\begin{aligned}
E\hat{g}(x) - g(x) &= EK_h(x - X_t) - g(x) \\
&\quad \text{(by identical distribution)} \\
&= \frac{1}{h}\int_a^b K\left(\frac{x-y}{h}\right) g(y)dy - g(x) \\
&= \int_{(a-x)/h}^{(b-x)/h} K(u)g(x+hu)du - g(x) \\
&\quad \text{(by change of variable } \frac{y-x}{h} = u) \\
&= \int_{-\lambda}^{1} K(u)g(x+hu)du - g(x) \\
&= g(x)\int_{-\lambda}^{1} K(u)du - g(x) \\
&\quad + h\int_{-\lambda}^{1} uK(u)g'(x+\tau hu)du \\
&\quad \text{(by the mean-value theorem, where } \tau \text{ lies in } (0,1)) \\
&= g(x)\left[\int_{-\lambda}^{1} K(u)dx - 1\right] + O(h). \\
&= O(1)
\end{aligned}$$

if $g(x) \geq \epsilon > 0$ for all $x \in [a, b]$ for any small but fixed constant ϵ.

Boundary problem of kernel estimators: If $x \in [a, a+h)$ or $(b-h, b]$, $E\hat{g}(x) - g(x)$ never vanishes to zero even if $h \to \infty$. This is due to the fact that there is no symmetric coverage of data in the boundary region $[a, a+h)$.

Question: What is the solution?

- **Trimming:**

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

  Remark: 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 and welfare economics (e.g., the low income population).

- **Using a boundary kernel:**

  To modify the kernel $K[(x - X_t)/h]$ such that it becomes location-dependent.

  A simple method (Hong and Li 2005):

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

  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. The idea is to modify the kernel function in the boundary region so that the integral of the kernel function is unity. Then the bias is $O(h^2)$ for all $x \in [a+h, b-h]$ and is $O(h)$ for $x \in [a, a+h)$ and $(b-h, b]$. 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.

  Another method: **The Jackknife kernel:** For x in the interior region $[a+h, b-h]$, use the standard 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 $\xi \in [0,1]$. 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 $x \in [a,b]$.

  Remark: The jackknife kernel formula in Härdle (1990, Section 4.4) is incorrect.

- **The Reflection method:**

  The reflection method is to construct the kernel density estimate based on 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 $K(\cdot)$ having support on $[-1,1]$, when x is away from the boundary, the second term will be zero. Hence, it only corrects the estimate in the boundary region. See Schuster (1985, *Communications in Statistics: Theory and Methods*) and Hall and Wehrly (1991, *Journal of American Statistical Association*).

  Question: What is the correct formula for the kernel density estimator when the support of X_t is $[a,b]$?

  Answer:

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

- **Transformation:** $Y_t = q(X_t)$, where $q(\cdot)$ is a given monotonic increasing function ranging 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 $\{Y_t\}_{t=1}^T$ which has infinite support.

  Question: Is there any free lunch?

- **Local polynomial fitting**

  This will be described later.

Question: We have deal with the bias of $\hat{g}(x)$. What is the variance of $\hat{g}(x)$?

For the time being, in order to simplify the analysis, we assume:

Assumption: $\{X_t\}$ is i.i.d.

Remark: This assumption simplies our analysis of computing the asymptotic variance of $\hat{g}(x)$. Later, we can relax the independence assumption for $\{X_t\}$ such that $\{X_t\}$ is a so-called α-mixing process. This will not change the asymptotic results for $\hat{g}(x)$.

Question: What is the α-mixing condition?

Put $Z_t = K_h(x - X_t) - EK_h(x - X_t)$. Then the variance

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

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

Remark: The variance of $\hat{g}(x)$ is proportional to $T^{-1}h^{-1}$, which is the approximate sample size for the observations in the interval $[x - h, x + h]$. It follows that the mean squared error (MSE) of $\hat{g}(x)$ is given by:

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

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

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

Remarks:

(i) For $\hat{g}(x) \to^p g(x)$, we need $Th \to \infty, h \to 0$, as $T \to \infty$.

(ii) It is always consistent for $g(x)$ but at a slower rate than $T^{-1/2}$. This means that a large sample is needed to obtain a reasonable estimate for $g(x)$.

(iii) Moreover, the bias depends on the smoothness of the unknown function $g(\cdot)$. If the second derivative has a sharp spike at the point x, then it is very difficult to obtain a good estimate $g(\cdot)$ at the point x.

Remark: Relative MSE:

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

if $g(x) > 0$.

Remark: It is very difficult to obtain a reasonable estimate of $g(x)$ in the sparse area where relatively few observations are available, or in the area where $g(\cdot)$ changes dramatically.

2.1.3 Optimal Choice of the Bandwidth

Remark: The choice of the optimal bandwidth can be obtained by minimizing $MSE[\hat{g}(x), g(x)]$:

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

The less smooth $g(x)$ is or the more sparse the observations are, the smaller the bandwidth h_0. This gives the optimal convergence rate for $\hat{g}(x)$:

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

Remark: The optimal bandwidth is unknown, because it depends on the unknown $g(x)$ and its second order derivative that we are interested in!

Question: How to obtain this optimal rate in practice?

Plug-in method: Obtain some initial preliminary estimators, say $\tilde{g}(x)$ and $\tilde{g}''(x)$ for $g(x)$ and $g''(x)$ and then plug them into the above formula. With such a data-dependent bandwidth, we obtain a new kernel estimator which has better statistical properties than an arbitrary choice of h.

Remark: Even if $\tilde{g}(x)$ and $\tilde{g}''(x)$ are not consistent for $g(x)$ and $g''(x)$, then $\hat{g}(x)$ is still consistent for $g(x)$ but not optimal.

2.1.4 Optimal Choice of the Kernel Function

Using the calculus of variation, it can be shown, as does in Epanechnikov (1969, *Theory of Probability and Its Applications*) that the optimal kernel that minimizes the MSE over a class of kernel functions is the so-called Epanechnikov kernel:

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

Remark: The choice of h is more important than the choice of $K(u)$. See also Priestley (1962).

Question: What happens if $\{X_t\}$ is serially dependent. Suppose $\{X_t\}$ is an α-mixing process.

Answer: Under suitable conditions on $\alpha(j)$, for exampple, $\alpha(j) \leq Cj^{-\beta}$ for $\beta > \frac{5}{2}$, we have the same MSE for $\hat{g}(x)$ as we have when $\{X_t\}$ is i.i.d.

2.2 Kernel Estimation of a Multivariate Density Function

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})?

Consider the kernel estimator

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

where

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

We first consider the bias. For an interior point x such that $x_i \in [a_i + h, b_i - h]$ for all

$i = 1, ..., d,$

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

Remark: $\sum_{i=1}^{d} f_{ii}(x)$ is called the Laplace of the function $f(x)$.

The variance of $\hat{f}(x)$

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

The MSE of $\hat{f}(x)$:

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

Remarks:

(i) Th^d is approximately the sample size for a d-dimensional subspace with each size equal to h.

(ii) The optimal MSE convergence rate of $\hat{f}(x)$ to $f(x)$ is $O_p(T^{-\frac{4}{4+d}})$ which can be obtained by setting

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

Thus, the MSE convergence rate is

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

The larger dimension d, the slower convergence of $\hat{f}(x)$. This is the so-called "**curse of dimensionality**".

Question: How to deal with the curse of dimensionality?

Reduction of dimensionality

- Assumption: Multiplicability conditions such as

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

- Assumption: Suppose $\{X_t\}$ is a Markov process:

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

Question: What happens if $\{X_t\}$ is serially dependent?

Answer: As long as the serial dependence of X_t on its past history is not too strong (for example, it satisfies a so-called strong mixing condition; see White (1999, *Asymptotic Theory for Econometricians*, 2nd Edition)), then the results established above continue to hold. This follows because the covariance terms $\text{cov}[K_h(x, X_t), K_h(x, X_s)]$ together are of smaller order in magnitude than $\text{var}[K_h(x, X_t)]$, due to the smoothing parameter.

Question: What is the α-mixing condition?

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 $j \to \infty$.

Remarks:

(i) Similar to ergodicity, the α-mixing condition is a notion 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.

(ii) α-mixing implies ergodicity. See White (*Asymptotic Theory for Econometricians*, 1999).

(iii) 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.

(iv) 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_0, X_j).

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

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

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

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

where $r = (1 - 1/p - 1/q)^{-1}$.

(ii) 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 $\alpha(j) \leq Cj^{-\beta}$ for some $C > 0$ and $\beta > 2$. Assume that $f_j(x,y)$ is bounded uniformly in (x,y) and in j. Then for $x \in [a,b]$,

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

Proof: Let $Z_t = K_h(x, X_t)$. Then by the stationarity of $\{X_t\}$, we have

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

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

$$\begin{aligned} \text{var}(Z_1) &= EK_h^2(x, X_1) - (EZ_1)^2 \\ &= h^{-1}g(x)D_K + o(h^{-1}). \end{aligned}$$

Thus, we only need to show

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

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,

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

Hence, we have

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

By taking $m(T) = h^{-2/\beta}$, 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.

Remark: The asymptotic variance of $\hat{g}(x)$ is the same as that under the i.i.d. assumption on $\{X_t\}$.

Question: Why?

Intuition [Hart (1996)]. Suppose the kernel $K(\cdot)$ has support on $[-1, 1]$. Then the kernel density estimator at the point x uses only the local data points with 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 $\{X_t\}_{=1}^{T}$ in the original sequence can be highly correlated, the dependence for the new series 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*)

Question: How to estimate $\hat{g}(x)$ using real data?

Applications of Density Estimation in Economics and Finance

- Ait-Sahalia (1996, *Review of Financial Studies*): Use the kernel-based marginal density estimator $\hat{g}(x)$ to test the adequacy of a diffusion model.

- 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.

- 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 (2003, *Working paper*): 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.

- de Matos, J.A. and M. Fernandes (2001, *Working paper*): How to test the Markov property for a time series process?

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

  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

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

How to check whether a stationary time series $\{X_t\}$ is a Markovian process? This requires to check whether

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

where $I_{t-1} = \{X_t, X_{t-1}, ...,\}$. **Question:** Why is this important?

Remark: Most continuous-time diffusion models are Markovian processes.

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

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

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

$$\{Z_t\} \sim i.i.d.U[0,1].$$

A potential Nonparametric Test:

Construct a nonparametric density estimator

$$\hat{f}(x|y) = \frac{\hat{f}(x,y)}{\hat{g}(y)},$$

where

$$\begin{aligned}\hat{f}(x,y) &= \frac{1}{(T-1)h^2}\sum_{t=1}^{T} K_h(x,X_t)K_h(y,X_{t-1}),\\ \hat{g}(y) &= \frac{1}{Th}\sum_{t=1}^{T} K_h(y,X_t).\end{aligned}$$

Define

$$\hat{Z}_t = \int_{-\infty}^{X_t} \hat{f}(x|X_{t-1})dx.$$

Then one can check if $\{\hat{Z}_t\}$ is approximately i.i.d.

We need to construct a test statistic and derive its asymptotic distribution. A related work is Hong and Li (2005). Note that it is important to deal with the possible impact of the sampling variation in $\hat{f}(\cdot)$.

3 Nonparametric Regression Estimation

Question: How to estimate a regression function $E(Y_t|X_t)$ using the sample $\{Y_t, X_t\}_{t=1}^T$?

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

Example 2: The conditional variance

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

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 a Markovian process. Then

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

This is the 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 we assume that $\{X_t\}$ is a Markovian process. Then

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

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

3.1 Kernel Regression Estimation

Assumption: Suppose $\{Y_t, X_t'\}'$ is an i.i.d. sequence such that $r(x) \equiv E(Y_t | X_t = x)$ exists and $r(x)$ is twice continuously differentiable.

Remark: We will relax the i.i.d. assumption to a dependent time series process at a later stage. In fact, to allow some mild serial dependence (e.g.., α-mixing) in $\{Y_t, X_t'\}'$ will not affect the asymptotic results derived under the i.i.d. assumption.

Question: How to estimate $r(x)$?

We can always write

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

where $E(\varepsilon_t | X_t) = 0$ a.s. and $\text{var}(\varepsilon_t | X_t) = \sigma^2(X_t)$ a.s. Note that we allow for conditional heteroskedasticity.

Define the so-called Nadaraya-Watson estimator

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

where

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

and, as before,

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

is the kernel estimator for density $g(x)$.

Alternatively, we can write

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

where the weighting function

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

satisfies

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

Geometric Interpretation:

Suppose the uniform kernel $K(u) = \frac{1}{2}\mathbf{1}(|u| \leq 1)$ is used. Then

$$\begin{aligned} \hat{r}(x) &= \frac{\sum_{t=1}^{T} Y_t \mathbf{1}(|X_t - x| \leq h)}{\sum_{t=1}^{T} \mathbf{1}(|X_t - x| \leq h)} \\ &= \text{the average of the } Y_t \text{ whose corresponding } X_t\text{'s fall into the interval } [x-h, x+h] \\ &= \text{the local sample mean.} \end{aligned}$$

Remark: More generally, we can assign different weights to different observations according to their distances to the location x. This will make sense because the observations closer to x will contain more information about $r(x)$. The use of $K(\cdot)$ is to assign different weights for observations.

Remark: Kernel regression is a special convolution filter used in engineering.

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

Observe

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

Here the second term is of a higher order term because $\hat{g}(x) - E\hat{g}(x) \to^p 0$ and $E\hat{g}(x) \to g(x)\int_{-1}^{1} K(u)du > 0$.

Remark: It can be shown that the second term is of a higher order term that vanishes faster than the first term (question: 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.

We first consider the numerator

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

For the variance component, we have

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

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 $X_t = x$.

On the other hand, for the denominator, we have

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

if $\int K(u)du = 1$. It follows that

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

Remark: The 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 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, we first write

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

For all interior points $x \in [a+h, b-h]$, we have

$$\begin{aligned}
E\hat{B}(x) &= E[r(X_t)K_h(x-X_t)] - r(x)E[K_h(x-X_t)] \\
&= \int r(z)K_h(x-z)g(z)dz - r(x)\int K_h(x-z)g(z)dz \\
(\text{define } m(z) &= r(z)g(z)) \\
&= \int_a^b m(z)K_h(x-z)dz - r(x)\int g(z)K_h(x-z)dz \\
&= \int_{(a-x)/h}^{(b-x)/h} m(x+hu)K(u)du \\
&\quad -r(x)\int_{(a-x)/h}^{(b-x)/h} g(x+hu)K(u)du \\
&= m(x)\int_{-1}^{1} K(u)du \\
&\quad +hm'(x)\int_{-1}^{1} uK(u)du \\
&\quad +\frac{1}{2}h^2m''(x)\int_{-1}^{1} u^2K(u)du[1+o(1)] \\
&\quad -r(x)g(x)\int_{-1}^{1} K(u)du \\
&\quad -hr(x)g'(x)\int_{-1}^{1} uK(u)udu \\
&\quad -\frac{1}{2}h^2r(x)g''(x)\int_{-1}^{1} u^2K(u)du][1+o(1)] \\
(\text{given } \int_{-1}^{1} uK(u)du &= 0) \\
&= \frac{1}{2}h^2\left[m''(x) - r(x)g''(x)\right]\int u^2K(u)du[1+o(1)] \\
&= \frac{1}{2}h^2[r''(x)g(x) + 2r'(x)g'(x)]C_K + o(h^2),
\end{aligned}$$

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

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

where we have made use of the fact that

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

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

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

Answer: Yes, because we still have

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

Question: What happens to the bias $E\hat{B}(x)$ if $x \in [a, a+h) \cup (b-h, b]$. Does $E\hat{B}(x) \to 0$ as $h \to 0$?

Answer: Yes, we still have $E\hat{B}(x)/E\hat{g}(x) = O(h)$ for x in the boundary region (say, $x = \tau h$ for $\tau \in [0, 1]$). This is different from the kernel density estimator $\hat{g}(x)$. However, it is slower than $O(h^2)$, the rate of the bias at the interior region.

Question: Why? This is because

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

Remarks: The boundary correction techniques are still useful to 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 $O(h^2)$.

We now show $\hat{B}(x) - E\hat{B}(x)$ is a higher order. Put

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

Then

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

is a higher order term.

It follows that

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

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

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

Remarks:

(i) The optimal choice of h is obtained by minimizing the MSE of $\hat{r}(x)$:

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

where

$$c^* = \left[\frac{D_K}{C_K} \frac{\sigma^2(x) g(x)}{[r''(x) + 2r'(x)g'(x)]^2} \right]^{\frac{1}{5}} T^{-\frac{1}{5}}.$$

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

(ii) The optimal choice of the kernel function: Like in the estimation of the probability density function, it is still the Epanechnikov kernel

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

(iii) 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?

Answer: 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 MSE of $\hat{r}(\cdot)$ is not the same as the optimal bandwidth h that minimizes the MSE of $\hat{r}^{(d)}(\cdot)$, where $d = 1, 2$. A larger bandwidth is needed to estimate the derivatives of $r(x)$.

3.2 Local Polynomial Estimator/Local Weighted Least Square Estimator

Interpretation for the Nadaraya-Watson Estimator

Consider the problem

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

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

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

We now consider the 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. The FOC is given by

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

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 the so-called local constant estimator. Therefore, the kernel regression estimator can be viewed as a locally weighted sample mean.

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

Question: Is there any gain by using a local polynomial estimator?

Local Polynomial Regression Estimator

References:

Suppose z is a point in a small neighborhood of x, and $r(z)$ is 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 neighborhood of x,

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

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

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

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

We thus consider the local minimization problem

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

where $\alpha = (\alpha_0, \alpha_1, ..., \alpha_p)'$ and $Z_t = [1, (X_t - x), ..., (X_t - x)^p]'$. The resulting local weighted least squares estimator

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

is the so-called local polynomial estimator of $r(z)$ for z near x. In particular, $\hat{\alpha}_0$ is an estimator for $r(x)$, and $\nu!\hat{\alpha}_\nu$ is an estimator for $r^{(\nu)}(x)$ where $0 < \nu \leq p$.

Question: Why?

The regression estimator at point x is then given by

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

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 derivative estimator at point x

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

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

Remarks:

(i) Local polynomial smoothing is very convenient for estimating the $r^{(j)}(x)$ simultaneously.

(ii) When $p = 0$, we have a local constant estimator, i.e., the Nadaraya-Watson estimator.

(iii) To compute the local polynomial estimator, one has to choose p, h and $K(\cdot)$. 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. 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 and the plug-in methods.

Matrix Expression

Put

$$\begin{aligned} Z_t &= [1, (X_t - x), (X_t - x)^2, ..., (X_t - x)^p]', \text{ a } (p+1) \times 1 \text{ regressor vector,} \\ W_t &= K_h(x - X_t) = h^{-1}K\left(\frac{x - X_t}{h}\right). \end{aligned}$$

Then

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

FOC:

$$\begin{aligned} \sum_{t=1}^{T} Z_t W_t(Y_t - Z_t'\hat{\alpha}) &= 0. \\ \sum_{t=1}^{T} Z_t W_t Y_t &= \left(\sum_{t=1}^{T} Z_t W_t Z_t'\right)\hat{\alpha} \end{aligned}$$

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 $W = diag(W_1, \cdots, W_T)$ is a $T \times T$ diagonal matrix, Z is a $T \times (p+1)$ matrix, and Y is a $T \times 1$ vector.

Remark: This is a local WLS (when $K(\cdot)$ has a bounded support on $[-1, 1]$)!

Question: What is the advantage of using the local polynomial approximation?

Asymptotic Properties of Local Polynomial Smoothing

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 $(\nu+1)$ position and zero elsewhere. Recalling $W_t = K_h(x - X_t) = h^{-1}K[(x - X_t)/h]$, we define

$$\hat{S}_j = \sum_{t=1}^{T} (X_t - x)^j K_h(X_t - x) = \sum_{t=1}^{T} (X_t - x)^j W_t,$$

and let

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

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

Then we have $\hat{\alpha} = \hat{S}^{-1}Z'WY$, and so

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

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

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

where $H = diag(1, h, ..., h^p)$ and $P(u) = (1, u, ..., u^p)'$ is a $(p+1)\times 1$ vector of a p-th order polynomial. Recall we will make change of variable $u = (X_t - x)/h$.

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

Lemma [Orthogonality]:

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

where $\delta_{\nu,q} = 1$ if $\nu = q$ and $\delta_{\nu,q} = 0$ otherwise.

Question: What is the intuition behind this orthonormality result?

Proof: Observing that $(X_t - x)^q = Z_t' e_{q+1}$, 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 the $(p+1)\times(p+1)$ matrix whose (i,j)th element is μ_{i+j-2}, where $\mu_j = \int_{-\infty}^{\infty} u^j K(u)du$. Then

$$S = \int P(u)K(u)P(u)'du.$$

Define the equivalent kernel by

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

Lemma [Equivalent Kernel]: Suppose $\{Y_t, X_t\}$ is a stationary α-mixing process with $\alpha(j) \le Cj^{-\beta}$ for $\beta > \frac{5}{2}$. Suppose that the marginal 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 that $K(\cdot)$ satisfies a Lipschitz condition. Then

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

where $a_T = [\ln(T)/Th]^{1/2} + h$. Moreover, the equivalent kernel $K_\nu^*(\cdot)$ satisfies the following moment condition

$$\int_{-\infty}^{\infty} u^q K_\nu^*(u)du = \delta_{\nu,q}, \text{ for } 0 \leq \nu, q \leq p.$$

Remark: The lemma implies that

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

Thus, the local polynomial estimator works like a kernel regression estimator with a known design density $g(x)$. This explains why the local polynomial estimator adapts to various design densities. In particular, it fits well even where $g'(x)$ is large. In the regions where $g'(x)$ is large, the standard Nadaraya-Watson kernel estimator cannot fit well, due to large biases.

Proof: We first consider the denominator $\hat{S} = [\hat{S}_{i+j-2}]_{(i,j)}$. Observe that

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

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

$$(Th^j)^{-1}\hat{S}_j = g(x)\mu_j + O_P(a_T),$$

where $O(h)$ in $a_T = [(h/T)^{1/2} \ln T + h]$ is contributed by the bias term in a first order Taylor series expansion. Recall $H = diag(1, h, ..., h^p)$. It follows that

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

or equivalently

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

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

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

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

The properties for the equivalent kernel $K^*_\nu(u)$ can be shown in the same way as the proof of the first Lemma. Observing $u^q = P(u)'e_{q+1}$, we have

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

This completes the proof of the second lemma.

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

We first write the v-th component of $\hat{\alpha}$,

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

For the first term, using $\hat{S} = Tg(x)HSH[1 + O(a_T)]$, which has been proven earlier, we can write

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

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

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

by the change of variable and the continuity of $\sigma^2(\cdot)$, where S^* is the $(p+1)\times(p+1)$ matrix with (i,j)-th element $\int u^{i+j-2}K^2(u)du$. Note that $S^* \neq S$. For S, the (i,j)-th element is $\int u^{i+j-2}K^2(u)du$.

It follows that the variance

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

Question: How to express $e'_{\nu+1}S^{-1}S^*S^{-1}e_{\nu+1}$ using the equivalent kernel $K^*_\nu(\cdot)$?
Answer: Recall $K^*_\nu(u) = e'_{\nu+1}P(u)K(u)$. We have

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

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

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

(Recall $\alpha_\nu = \nu! r^{(\nu)}(x)$.) Put the reminder

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

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

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

by Chebyshev's inequality.

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

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

which is a higher order term (**Question**: How to show this under the i.i.d. assumption?). Thus, the bias is determined by

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

where C is a $(p+1) \times 1$ vector with the i-th element $\int u^{p+2-i} K(u) du$.

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

Answer: Recall $K_\nu^*(u) = e'_{\nu+1} S^{-1} P(u) K(u)$, we have

$$\int_{-1}^{1} u^{p+1} K_\nu^*(u)du = e'_{\nu+1} S^{-1} \int_{-1}^{1} u^{p+1} P(u) K(u)du = e'_{\nu+1} S^{-1} C.$$

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

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

Remarks:

(i) The local WLS can consistently estimate the Taylor series expansion coefficients:

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

(ii) By minimizing the MSE, the optimal convergence rate can be achieved by choosing the bandwidth

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

The optimal bandwidth h^* does not depend on the order of the derivative ν. Of course, the proportionality still depends on ν.

(iii) The intuitive idea of local polynomial smoothing in economics can be dated back to Nerlove (1966), where he use a piecewise linear regression to estimate a nonlinear cost function for the electricity industry. Also see White (1980, International Economic Review) for a related discussion.

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

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

Boundary Behavior of the Local Polynomial Estimator

Question: The above results 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 a reasoning analogous to what we have done above, we can obtain

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

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 $-\infty$ to τ. For example, S_τ is a $(p+1)\times(p+1)$ matrix, with (i,j)-th element equal to

$$\mu_{i+j-2,\tau} = \int_{-\tau}^{\infty} u^{i+j-2}K(u)du.$$

Interestingly, the biases of $\hat{\alpha}_\nu(x)$ are of the same order of magnitude no matter 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). Thus, the local polynomial estimator automatically adopts to the boundary region and does not suffer from the boundary bias problem of the standard kernel method.

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

Answer: The key is the joint use of the local intercept and local slope (for a local linear smoother). The latter can handle asymmetric behaviors such as those in the boundary regions.

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

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

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

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

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

Proof: Similarly to the derivation of MSE for the local polynomial in the interior point.

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

Remarks:

(i) It avoids the boundary problem in regression estimation.

(ii) It has a smaller bias term for the regression estimator when the marginal density $f(x)$ of X_t is a large derivative (i.e., when $f'(x)$ is large), and consequently is more efficient than traditional kernel estimators (the Nadaraya-Watson estimator).

Applications of Regression Smoothing in Economics and Finance:

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

$$\begin{aligned} G_t &= G(X_t, P_t, \tau_t, r_{t,T}) \\ &= \exp[-r_{t,t}(T-t)] \int Y(P_t, X_t) f^*_t(P_T; T) dP_t, \end{aligned}$$

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

  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.

- Ait-Sahalia (1996), Stanton (1997) and Chapman and Pearson (1999): Use non-parametric kernel estimators $\hat{r}(X_{t-1})$ to estimate $E(X_t|X_{t-1})$, where X_t is the spot interest rate, and investigate whether the the drift function $\mu(X_t)$ in the diffusion model

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

  is nonlinear.

Questions: What are potential topics for research in time series econometrics?

Example 1: Asset Pricing Models with Time-Varying β coefficients:

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

where

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

Constant beta vs. nonconstant beta? From the Euler equation,

$$\begin{aligned} \alpha_i &= \alpha_i(I_{t-1}), \\ \beta_i &= \beta_i(I_{t-1}) \end{aligned}$$

are possibly time-varying coefficients. Suppose $\alpha_i = \alpha_i(Z_t)$ and $\beta_i = \beta_i(Z_t)$, where Z_t is some state variable or vector in I_{t-1}. Then one can estimate $\alpha_i(\cdot)$ and $\beta_i(\cdot)$ by solving the problem

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

Question: What is the economic rationale that α and β are time-varying?

Reference: Kevin Wang (2002, *Journal of Finance*).

Example 2: Time-varying risk aversion parameter and risk premium puzzles.

Consider the problem

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

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 problem is

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

where $\frac{P_{t+1}+D_{t+1}}{P_t}$ 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 economic agent.

Suppose the utility function of the 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.

With the CRRA utility function, the Euler equation becomes

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

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

Question: What is the "risk premium puzzle"?

Answer: The risk premium puzzle exists because the excess of stock returns of stock returns over returns on investments in bills or bonds is larger than can 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, 145-161).

A possible solution: Assume both β and γ are time-varying: $\beta_t=\beta(I_{t-1})$ and $\gamma_t=\gamma(I_{t-1})$, where I_{t-1} is the information set. More specifically, we can assume $\beta_t=\beta(Z_t)$ and $\gamma_t=\gamma(Z_t)$, for some unknown smooth function $\beta(\cdot)$ and $\gamma(\cdot)$, where $Z_t\in I_{t-1}$ is a state vector that is expected to affect both β and γ. These time-varying functions can reveal very useful information about how the risk attitude of the economic agent changes with the state variable or vector Z_t.

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

Recall that the Euler equation is a conditional mean specification (i.e., regression analysis). Therefore, we can estimate $\beta(\cdot)$ and $\gamma(\cdot)$ using the local polynomial method:

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

where $\beta(x)$ and $\gamma(x)$ are some low-order polynomial estimators.

Example 3: Functional-Coefficient Regression Models for Nonlinear Time Series:

$$E(X_t|I_{t-1})=\sum_{j=1}^{d}a_j(X_{t-d})X_{t-j}.$$

Example 4: Volatility Smile and Correct Derivative Pricing.

Black-Scholes (1973) formula for the price of an European call option

Assumptions:

(A1) $dS_t = \mu S_t dt + \sigma S_t dW_t$, 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: payoff function

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

where K is the strike price.

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

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

where $t = T - \tau_t$, and

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

Volatility smile: $\sigma_t^2 = \sigma^2(K_t, S_t, r_t, \tau_t, \pi_t)$ is convex in strike price K_t if the pricing is incorrect. If the pricing formula is correct, then 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.

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

4 Nonparametric Estimation of Time-Varying Parameter Models

Example 1 [Estimation of a Slow-Varying Time Trend Function]

Suppose a time series process

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

where $f(\cdot)$ is a smooth but unknown time-trend function and $\{X_t\}$ is a stationary process with $E(X_t) = 0$.

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

We can separate the smooth trend from the noisy stochastic error with smoothing techniques.

References:

Hall and Hart (1990),
Johnstone and Silverman (1997),
Robinson (1997)

Suppose $f(\cdot)$ is continuously differentiable on [0,1] up to order p, and we are interested in estimating the function $f(t_0/T)$ at t_0. We consider the local polynomial smoothing by solving the problem

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

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

$$\begin{aligned}
Z_t &= \left[1, \left(\frac{t-t_0}{T} \right), ..., \left(\frac{t-t_0}{T} \right)^p \right]', \\
W_t &= K_h \left(\frac{t-t_0}{T} \right) = \frac{1}{h} K \left(\frac{t-t_0}{Th} \right).
\end{aligned}$$

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

$$\hat{\alpha}_\nu = e_{\nu+1}' \hat{\alpha},$$

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

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

Put $\hat{S} = Z'WZ'$. We first decompose

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

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

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

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

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

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

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

where S is a $(p+1)\times(p+1)$ matrix with (i,j) element $\int_{-1}^{1} u^{i+j-2}K(u)du$. Also, for each given j, by approximating the discrete sum with a continuous integral, we have

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

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

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

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

It follows that

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

Remark: Unlike the estimator for the regression function $r(X_t)$, the asymptotic variance of $\hat{f}(t_0/T)$ depends on the serial dependence in $\{X_t\}$. In other words, whether $\{X_t\}$ is i.i.d. has important impact on the variance of $\hat{\alpha}_\nu = \hat{f}^{(\nu)}(t_0/T)$.

Question: Why?

Answer: The local polynomial estimator for $f(t_0/T)$ is based on the observations in the local interval $[t_0 - hT, t_0 + hT]$. These observations maintain the same pattern of serial dependence as the original data. As a result, the estimator depends on the serial dependence of $\{X_t\}$.

Next, for the bias, using the Taylor series expansion

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

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

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

where C is a $(p+1) \times 1$ vector with the i-th element being $\int_{-1}^{1} u^{p+2-i} K(u) du$. Here, we have used a continuous integral to approximate a discrete sum:

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

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

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

Asymptotic Distribution

Example 2 [Estimation of Locally Stationary Processes]

Question: How to model smooth time changes in a system such as an economic system?

$$Y_t = X_t'\alpha + \varepsilon_t$$

We consider a locally stationary process

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

where Y_t is a scalar, X_t is a $(d+1)\times 1$ random vector, and $\alpha(t/T)$ is a $(d+1)\times 1$ smooth function of t.

Question: Why smooth changes?
Answer:

(i) Structral changes/breaks are rather a rule than an exception, due to advances in technology, changes in preferences, and institutional changes in the economic system.

(ii) 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, and it takes time for markets to reach some consensus due to heterogeous beliefs.

(iii) Even if individual agents can respond immediately to sudden changes, the aggregated economic variables (such as consumption) over many individuals will become smooth.

Alfredo Marshall: Economic changes are evolutionary.

Question: How to estimate the changing coefficients $\alpha(t/T)$?

Remark: This model is potentially useful for macroeconomic applications and for long time series data.

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 $(d+1)(p+1)\times 1$ vector. Then we consider

$$\begin{aligned}
\sum_{t=1}^{T}\left[Y_t - \sum_{j=0}^{d}\alpha_{jt}'X_{jt}\right]^2 &= \sum_{t=1}^{T}\left[Y_t - \sum_{j=0}^{d}\alpha_{0j}'Z_tX_{jt}\right]^2 \\
&= \sum_{t=1}^{T}\left[Y_t - \alpha'(Z_t\otimes X_t)\right]^2 \\
&= \sum_{t=1}^{T}\left[Y_t - \alpha' Q_t\right]^2,
\end{aligned}$$

where $\alpha = (\alpha_0', \alpha_1', ..., \alpha_p')'$ is a $(d+1)(p+1)\times 1$ vector, α_j is a $d\times 1$ coefficient vector for $(\frac{t-t_0}{T})^j X_t$.

The local polynomial estimator

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

The estimator for $\alpha(t_0/T)$ is then given by

$$\hat{\alpha}_0 = (I_d \otimes e_{\nu+1})'\hat{\alpha}.$$

By plotting the $\hat{\alpha}_0$ as a function of t_0, we can examine whether the coefficient α is time-varying.

5 Nonparametric Kernel Method in Frequency Domain

Questions: Given $\{X_t\}_{t=1}^T$,

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

5.1 Periodogram and Motivation

Parametric Approach

For simplicity of analysis, we assume $\mu \equiv E(X_t) = 0$ and we know it. Then 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.)

Recall the power spectral density

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

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

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

where $\gamma(0) = \text{var}(X_t)$. A spectral estimator is

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

For an MA(1) process, the spectral density

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

A 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

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

A spectral 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 parameter estimators, and

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

where

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

with $\hat{\varepsilon}_t = 0$ if $t \leq 0$.

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

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

Remark: $\hat{I}_T(\omega)$ is the so-called periodogram of $\{X_t\}_{t=1}^T$. It is the squared modulus of the discrete Fourier transform of data $\{X_t\}_{t=1}^T$. Please check that the last equality holds.

Remark: Unfortunately, this estimator is not consistent for $h(\omega)$. Why?

For example, consider the simplest case when $\{X_t\}$ is i.i.d. Then we have $h(\omega) = \frac{1}{2\pi}\gamma(0)$, and

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

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

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

It follows that

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

Remark: The variance $\text{var}[\hat{h}(\omega)]$ never decays to 0.

Why? Too many estimated coefficients $\{\hat{\gamma}(j)\}_{j=0}^{T-1}$! There are T estimated coefficients.

Alternative Explanation: Integrated MSE (IMSE)

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

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

Note that $E\hat{\gamma}(j) = T^{-1} \sum_{t=|j|+1}^{T} E(X_t X_{t-|j|}) = (1 - |j|/T)\gamma(j)$, so we have the bias square

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

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

What is the variance?

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

because $E[\hat{\gamma}(j) - E\hat{\gamma}(j)]^2 = CT^{-1}$ under certain regularity conditions, for some $C > 0$.

Remark: The variance of the periodogram $\hat{I}_T(\omega)$ does not vanish.

5.2 Kernel Spectral Estimation

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

Truncation

In response to the fact that the periodogram is not consistent for $h(\omega)$ 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 $p \to \infty, p/T \to 0$.

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

We can use a weighted estimator. Consider

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

where $k(\cdot)$ is a kernel function or 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). The Bartlett kernel-based spectral density estimator at frequency zero is always positive semi-definite.

Remark: This is used in the so-called Newey-West (1987, *Econometrica,* 1994, *Review of Economic Studies*) variance-covariance estimator.

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

Answer: This reduces the variance of $\hat{h}(\omega)$.

Intuition: 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 impact 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 Daniel kernel

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

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

Remark: Now p is no longer a lag order but a smoothing parameter.

Assumption on $k(\cdot)$: The 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) $\int k^2(z)dz < \infty$, and (iv) there exists a positive real number q such that

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

Remark: For the Bartlett kernel, $q = 1$. For the Daniell kernel, $q = 2$.

Examples of $k(\cdot)$:

1. Bartlett kernel
$$k(z) = (1 - |z|)1(|z| \leq 1).$$
Its Fourier transform
$$K(u) = \frac{1}{2\pi} \left[\frac{\sin(u/2)}{u/2} \right]^2.$$

2. Daniell kernel
$$k(z) = \frac{\sin(\pi z)}{\pi z}.$$
Its Fourier transform
$$K(u) = \frac{1}{2\pi} 1(|u| \leq \pi).$$

3. Parzen kernel
$$k(z) = \begin{cases} 1 - 6z^2 + 6|z|^3 & |z| \leq \frac{1}{2} \\ 2(1 - |z|)^3 & \frac{1}{2} \leq |z| < 1. \\ 0 & \text{otherwise.} \end{cases}$$
Its Fourier transform
$$K(u) = \frac{3}{8\pi} \left[\frac{\sin(u/4)}{u/4} \right]^4.$$

4. Quadratic-Spectral kernel (Priestley)
$$k(z) = \frac{3}{(\pi z)^2} \left[\frac{\sin \pi z}{\pi z} - \cos(\pi z) \right].$$
Its Fourier transform
$$K(u) = \frac{3}{4\pi} [1 - (u/\pi)^2] 1(|u| \leq \pi).$$

5. Truncated kernel
$$k(z) = 1(|z| \leq 1).$$
Its Fourier transform
$$K(u) = \frac{1}{\pi} \frac{\sin u}{u}.$$

Remark: The kernel function $k(\cdot)$ used for spectral density estimation is the Fourier transform of a kernel function $K(\cdot)$ used in probability density/regression function estimation.

5.3 Consistency of Kernel Spectral Estimators

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

We consider the integrated MSE criterion

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

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,

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

as $p \to \infty$, where, as defined earlier, $k_q = \lim[1 - k(z)]/|z|^q$, 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 wen $\gamma(j)$ 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 $\sum_{j=1-T}^{T-1} |j|\, |\gamma(j)| < \infty$.

Similarly, the last term

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

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

Thus, suppose $T^{-1} = o(p^{-q})$, which can be satisfied by choosing a suitable bandwidth p, we have

$$E\hat{h}(\omega) - h(\omega) = -p^{-q} k_q h^{(q)}(\omega) + o(p^{-q})$$

and

$$\begin{aligned} & \int_{-\pi}^{\pi} \left[E\hat{h}(\omega) - h(\omega) \right]^2 d\omega \\ = \; & p^{-2q} k_q^2 \int \left[h^{(q)}(\omega) \right]^2 d\omega + o(p^{-2q}) \\ = \; & p^{-2q} k_q^2 \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} |j|^{2q} \gamma^2(j) + o(p^{-2q}). \end{aligned}$$

Remarks:

(i) If $h^{(q)}(\omega) > 0$, then the bias is always negative. In other words, the kernel method always underestimate spectral peaks.

(ii) There is no boundary bias problem, because $\hat{h}(\cdot)$ is a symmetric periodic function.

Next, for the variance of $\hat{h}(\omega)$, we have

$$\begin{aligned} & E\int[\hat{h}(\omega) - E\hat{h}(\omega)]^2 d\omega \\ = & \left\{\frac{1}{2\pi}\sum_{j=1-T}^{T-1} k^2(j/p)E[\hat{\gamma}(j) - E\hat{\gamma}(j)]^2\right\} \\ = & \frac{p}{T}\int_{-\pi}^{\pi} h^2(\omega)d\omega \int_{-\infty}^{\infty} k^2(z)dz[1+o(1)]. \end{aligned}$$

$$\begin{aligned} E[\hat{\gamma}(j) - E\hat{\gamma}(j)]^2 &= \text{var}[\hat{\gamma}(j)] \\ &\tilde{} \frac{1}{T}\int h^2(\omega)d\omega \\ &= \frac{1}{T}\Sigma_{j=-\infty}^{\infty}\gamma^2(j) \end{aligned}$$

Here, we have used the identity (see Priestley, 1981, p.) that

$$\begin{aligned} & \text{var}[\hat{\gamma}(j)] \\ = & T^{-1}\sum_{m=1-(T-j)}^{T-j-1}\left[1 - \frac{|m|+j}{T}\right]\left[\gamma^2(m) + \gamma(m+j)\gamma(m-j) + \kappa_4(m,j,m+j)\right], \end{aligned}$$

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 covariance as $\{X_t\}$. It follows that

$$\begin{aligned} & \frac{1}{(2\pi)}\sum_{j=1-T}^{T-1} k^2(j/p)\text{var}[\hat{\gamma}(j)] \\ = & \frac{1}{(2\pi)}\sum_{j=1-T}^{T-1} k^2(j/p)T^{-1}\sum_{m=1-(T-j)}^{T-j-1}\left[1 - \frac{|m|+j}{T}\right]\gamma^2(m) \\ & + \frac{1}{(2\pi)}\sum_{j=1-T}^{T-1} k^2(j/p)T^{-1}\sum_{m=1-(T-j)}^{T-j-1}\left[1 - \frac{|m|+j}{T}\right]\gamma(m+j)\gamma(m-j) \\ & + \frac{1}{(2\pi)}\sum_{j=1-T}^{T-1} k^2(j/p)T^{-1}\sum_{m=1-(T-j)}^{T-j-1}\left[1 - \frac{|m|+j}{T}\right]\kappa_4(m,j,m+j) \\ = & \hat{V}_1 + \hat{V}_2 + \hat{V}_3, \end{aligned}$$

where

$$\begin{aligned}\hat{V}_1 &= \frac{p}{T}\frac{1}{2\pi}\sum_{m=-\infty}^{\infty}\gamma^2(m)\left[\frac{1}{p}\sum_{j=1-T}^{T-1}k^2(j/p)\right] \\ &= \frac{p}{T}\int_{-\pi}^{\pi}h^2(\omega)d\omega\int_{-\infty}^{\infty}k^2(z)dz[1+o(1)],\end{aligned}$$

$$|\hat{V}_2| \leq \frac{1}{T}\sum_{j=-\infty}^{\infty}|\gamma(j)|\sum_{m=-\infty}^{\infty}|\gamma(m)| = O(T^{-1})$$

if $\sum_{m=-\infty}^{\infty}|\gamma(m)|<\infty$, and finally, for the last term,

$$|\hat{V}_3| \leq \frac{1}{T}\sum_{j=-\infty}^{\infty}\sum_{l=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}|\kappa_4(i,j,k)| = O(T^{-1}).$$

It follows that the IMSE

$$\begin{aligned}IMSE(\hat{h},h) &= \frac{p}{T}\int_{-\pi}^{\pi}h^2(\omega)d\omega\int_{-\infty}^{\infty}k^2(z)dz \\ &\quad +p^{-2q}k_q^2\int_{-\pi}^{\pi}\left[h^{(q)}(\omega)\right]^2d\omega \\ &\quad +o(p/T+p^{-2q}) \\ &= \frac{p}{T}\frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma^2(j)\int_{-\infty}^{\infty}k^2(z)dz \\ &\quad +p^{-2q}k_q^2\frac{1}{2\pi}\sum_{j=-\infty}^{\infty}|j|^{2q}\gamma^2(j) \\ &\quad +o(T^{-1}p+p^{-2q}) \\ &= O(p/T+p^{-2q}).\end{aligned}$$

Remark :

(i) $\hat{h}(\omega)$ is consistent for $h(\omega)$ for any $\omega\in[-\pi,\pi]$ if $p/T\to 0, p\to\infty$.

(ii)The optimal bandwidth

$$\begin{aligned}p_0 &= \left[\frac{2qk_q^2}{\int k^2dz}\frac{\int[h^{(q)}(\omega)]^2d\omega}{\int h^2(\omega)d\omega}\right]^{\frac{1}{2q+1}}T^{\frac{1}{2q+1}} \\ &= c_0T^{\frac{1}{2q+1}}.\end{aligned}$$

With this rate for p, the optimal convergence rate for $h(\omega)$ is $IMSE(\hat{h},h)\propto T^{-\frac{2q}{2q+1}}$.

This optimal bandwidth is unknown. Again, the plug-in method can be used.

(iii) The optimal kernel is the Quadratic-Spectral kernel

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

Note that the Fourier transform of the QS kernel is the Epanechnikov kernel

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

Question: What is the relationship between $K(u)$ and $k(z)$? They are Fourier transforms of each other:

$$\begin{aligned} K(u) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} k(z)e^{-izu}dz, \\ k(z) &= \int_{-\infty}^{\infty} K(u)e^{iuz}du. \end{aligned}$$

(i) $k(0) = 1$ is equivalent to $\int_{-\infty}^{\infty} K(u)du = 1$.
(ii) $\int_{-\infty}^{\infty} k^2(z)dz = 2\pi\int_{-\infty}^{\infty} K^2(u)du$. (How to prove this?)
(iii) When $q = 2$, $k_2 \equiv \lim_{z\to 0}\frac{1-k(z)}{z^2} = -\frac{1}{2}k''(0) = \frac{1}{2}\int_{-\infty}^{\infty} u^2K(u)du$.
(iv) $k(z)$ symmetric, then $K(u)$ is symmetric. So $\int_{-\infty}^{\infty} uK(u)du = 0$.

Question: Is there any equivalent expression for $\hat{h}(\omega)$ using $K(u)$?

Answer: Yes, recall

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

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

$$\begin{aligned} \hat{h}(\omega) &= \frac{1}{2\pi}\sum_{j=1-T}^{T-1} k(j/p)\hat{\gamma}(j)e^{-ij\omega} \\ &= \int_{-\pi}^{\pi} \hat{I}(\lambda)W_T(\omega - \lambda)d\lambda \\ &= \int_{-\pi}^{\pi} \hat{I}(\lambda)pK[p(\omega - \lambda)]d\lambda \\ &= \int_{-\pi}^{\pi} \hat{I}(\lambda)\frac{1}{h}K\left(\frac{\omega - \lambda}{h}\right)d\lambda, \end{aligned}$$

where $h = p^{-1}$, $\hat{I}(\lambda)$ is the periodogram,

$$\hat{I}(\lambda) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T} X_t e^{it\omega}\right|^2 = \frac{1}{2\pi}\sum_{j=1-T}^{T-1}\hat{\gamma}(j)e^{-ij\omega}$$

which is the discrete Fourier transform of $\hat{\gamma}(j)$, and

$$\begin{aligned} W_T(\lambda) &= \frac{1}{2\pi} \sum_{j=-(T-1)}^{T-1} k(j/p) e^{-ij\lambda} \\ &= p \left[\frac{1}{2\pi p} \sum_{j=-\infty}^{\infty} k(j/p) e^{-i(j/p)p\lambda} \right] \\ &= p \sum_{j=-\infty}^{\infty} K[p(\lambda + 2\pi j)] \\ &\sim pK(p\lambda), \end{aligned}$$

which is the discrete Fourier transform of $k(\cdot)$.

Remark: The periodogram $\hat{I}(\lambda)$ is the discrete Fourier transform of the observed data.

Question: When

$$p \sum_{j=-\infty}^{\infty} K[p(\lambda + 2\pi j)] = pK(p\lambda), \ \lambda \in [-\pi, \pi]?$$

Answer: When $K(\cdot)$ has bounded support on $[-\pi, \pi]$ and p is large (then the terms with $j \neq 0$ will vanish).

Remark : $p = h^{-1}$, the inverse of the bandwidth.

Remark : $\hat{I}(\lambda)$ is a (discrete) Fourier transform of $\hat{\gamma}(j)$, and $W_T(\lambda)$ 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.

Remark: $W_T(\lambda)$ plays a role of local weighting and smoothing.

Geometric interpretation of $\hat{h}(\omega) = \int_{-\pi}^{\pi} \hat{I}(\lambda) pK[p(\omega - \lambda)] d\lambda$:

Remark: Downward bias:

$$E\hat{h}(\omega) - h(\omega) = -p^{-q} k_q h^{(q)}(\omega) + o(p^{-q}).$$

This can be large when there is a spectral speak at frequency ω. What is the alternative estimation method?

Granger (1966): 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?

(i) Pre-whitening.

Tukey (1957), Andrews and Monahan (1992)

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 $\{u_t\}$ will have weaker serial dependence and

$$u_t = \Psi(L) X_t.$$

$$h_u(\omega) = |\Psi(e^{-i\omega})|^2 h_X(\omega).$$

Thus,

$$h_X(\omega) = |\Psi(e^{-i\omega})|^{-2} h_u(\omega).$$

Remarks:

(i) We first run a prewhitening regression, and obtain $\{\hat{\Psi}_j\}_{j=1}^m$.

(ii) Then use the kernel method to estimate $h_u(\omega)$ using the prewhitening residual $\{\hat{u}_t\}$.

(iii) Obtain $\hat{h}(\omega) = |\hat{\Psi}(e^{-i\omega})|^{-2} \hat{h}_u(\omega)$. This is called "recoloring".

The spectral density $h_u(\omega)$ is easier to estimate because it is "flatter" than $h(\omega)$.

Remark: The bias is reduced substantially but the variance is increased at the same time. As a consequence, MSE may be larger than that without using prewhitening.

(ii) Logarithmic transformation

Put $\lambda_k = 2\pi k/T$ for $k = 0, ..., [\frac{T-1}{2}]$. This is the so-called Fourier frequency. Then

$$\hat{I}_T(\lambda_k) = f(\lambda_k) V_k + R_k,$$

where

$$V_k = \frac{1}{2\pi T} \left| \sum_{t=1}^{T} \varepsilon_t e^{it\lambda_k} \right|^2$$

is the periodogram of an innovation sequence $\{\varepsilon_t\}_{t=1}^T$, and R_k is an asymptotically negligible term. For $0 < k < [\frac{T-1}{2}]$.

(iii) Wavelet analysis

References:

Härdle, W., G. Kerkyacharian, D. Picard and A. Tsybakov (1998), *Wavelets, Approximation and Statistical Applications,* Lecture Notes in Statistics Volume 129.

Hong, Y. and D. Kao (2004, *Econometrica*), Hong and Lee (2001, *Econometric Theory*), Lee and Hong (2001, *Econometric Theory*)

Question: How to estimate the generalized spectrum $f(\omega, u, v)$?

Kernel Method

$$\hat{f}(\omega, u, v) = \frac{1}{2\pi} \sum_{j=1-T}^{T-1} (1 - |j|/T)^{1/2} k(j/p) \hat{\sigma}_j(u, v) e^{-ij\omega},$$

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 $(X_t, X_{t-|j|})$.

Question: Why an additional factor $(1 - |j|/T)$?

Answer: It is introduced to improve the finite sample performance because $\hat{\varphi}_j(u, v)$ is based on the normalization $T - |j|$ instead of T.

Question: How to estimate the bispectrum $b(\omega_1, \omega_2)$?

$$\hat{b}(\omega_1, \omega_2) = \frac{1}{(2\pi)^2} \sum_{j=1-T}^{T-1} \sum_{l=1-T}^{T-1} k(j/p) k(l/p) k[(j-l)/p] \hat{C}(0, j, l) e^{ij\omega_1 + il\omega_2}.$$

Question: How to find the variance and the bias of $\hat{b}(\omega_1, \omega_2)$?

Answer: Suppose $k_2 = \lim \frac{1-k(z)}{|z|^2} \in (0, \infty)$. Then the bias

$$E\left[\hat{b}(\omega_1, \omega_2)\right] - b(\omega_1, \omega_2) = -\frac{1}{2}\frac{k_2}{p^2} D^{(2)}(\omega_1, \omega_2) + O(p^{-3})$$

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.

For the variance,

$$\text{var}\left[\hat{b}(\omega_1, \omega_2)\right] = \frac{p^2}{T}\frac{V}{2\pi} h(\omega_1) h(\omega_2) h(\omega_1 + \omega_2) \left[1 + o(1)\right],$$

where

$$V = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} k^2(u) k^2(v) k^2(u - v) du dv.$$

See Brillinger and Rosenblatt (1967a) for the derivation of the variance.

Remark: $\hat{b}(\omega_1, \omega_2)$ is consistent for $b(\omega_1, \omega_2)$ if $p \to \infty, p^2/T \to 0$ as $T \to \infty$.

Exercise 6

1. What are the main advantages of nonparametric smoothing methods in time series econometrics? Why has nonparametric methods become popular in the recent years?

2. What is the boundary problem for the kernel smoothing method? How can one alleviate this boundary problem?

3. What is the curse of dimensionality associated with nonparametric smoothing?

4. Why can the local linear smoother automatically solve for the boundary bias problem in nonparametric regression estimation?

5. Suppose $\{X_t\}_{t=1}^T$ is an i.i.d. 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.

(a) For $x \in [a + h, b - h]$, derive the asymptotic bias expression for $E\hat{g}(x) - g(x)$.

(b) For $x \in [a+h, b-h]$, derive the asymptotic variance expression for $\text{var}(\hat{g}(x)) = E[\hat{g}(x) - E\hat{g}(x)]^2$.

(c) Find the asymptotic expression for the mean squared error MSE $E[\hat{g}(x)-g(x)]^2$.

(d) Derive the optimal bandwidth h^* that maximizes the asymptotic MSE of $\hat{g}(x)$.

(e) What is the asymptotic MSE when evaluated at the optimal bandwidth h^*.

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 (a)–(e) in Question #5 again.

7. In the setup of Question #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 $\rho \in [0, 1)$.

(a) Show that $\sup_{x\in[a,a+h]} |E\hat{g}(x) - g(x)|$ never vanishes to zero as $h \to \infty$.

(b) There are several approaches to deal with the boundary bias problem in (a). 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 Question #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 $h \to 0$.

8. 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 (a) $x \in [a, a+h)$; (b) $x \in [ah, b-h]$.

9. Suppose a data generating process is given by

$$Y_t = 1 + X_t - 0.25X_t^2 + \varepsilon_t, \qquad t = 1, ..., T,$$

where $\{X_t\} \sim$ i.i.d.U$[0, 2\sqrt{3}]$, $\{\varepsilon_t\} \sim$ i.i.d.N(0,1), and $\{X_t\}$ and $\{\varepsilon_t\}$ are mutually independent.

(a) 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 $r(x) = E(Y_t|X_t = x)$.

(b) 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.

(c) 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 (b). Plot the estimator for $r(x)$ on the xy-plane.

10. Again, in the setup of Question #1, 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.

(a) 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

$$\bar{g}(x) = \frac{1}{T} \sum_{t=1}^{T} h^{-1} K_\rho^J \left(\frac{x - X_t}{h} \right).$$

(b) Find the expression for β in terms of $\omega_K(\cdot, \rho)$ and α such that $\sup_{x \in [a, a+h)} |E\bar{g}(x) - g(x)| = O(h^2)$.

(c) 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 $x \in (b - h, b]$.

EXERCISE 6

1. Suppose $\{X_t\}_{t=1}^T$ is an i.i.d. 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.

(a) For $x \in [a + h, b - h]$, derive the asymptotic bias expression for $E\hat{g}(x) - g(x)$.

(b) For $x \in [a+h, b-h]$, derive the asymptotic variance expression for $\text{var}(\hat{g}(x)) = E[\hat{g}(x) - E\hat{g}(x)]^2$.

(c) Find the asymptotic expression for the mean squared error MSE $E[\hat{g}(x)-g(x)]^2$.

(d) Derive the optimal bandwidth h^* that maximizes the asymptotic MSE of $\hat{g}(x)$.

(e) What is the asymptotic MSE when evaluated at the optimal bandwidth h^*.

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 (a)–(e) in Question #1 again.

2. In the setup of Question #1, 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 $\rho \in [0, 1)$.

(a) Show that $\sup_{x\in[a,a+h]} |E\hat{g}(x) - g(x)|$ never vanishes to zero as $h \to \infty$.

(b) There are several approaches to deal with the boundary bias problem in (a). 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 Question #1 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 $h \to 0$.

3. 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 (a) $x \in [a, a+h)$; (b) $x \in [ah, b-h]$.

4. Suppose a data generating process is given by

$$Y_t = 1 + X_t - 0.25X_t^2 + \varepsilon_t, \qquad t = 1, ..., T,$$

where $\{X_t\} \sim$ i.i.d.U$[0, 2\sqrt{3}]$, $\{\varepsilon_t\} \sim$ i.i.d.N(0,1), and $\{X_t\}$ and $\{\varepsilon_t\}$ are mutually independent.

(a) 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 $r(x) = E(Y_t|X_t = x)$.

(b) 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.

(c) 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 (b). Plot the estimator for $r(x)$ on the xy-plane.

5. 6. Again, in the setup of Question #1, 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.

(a) For $x = a + \rho h \in [a, a+h)$, we consider a $\bar{g}(x) = \hat{g}(x; h) + \beta\left[\hat{g}(x; h) - \hat{g}(x; \alpha h)\right]$, 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

$$\bar{g}(x) = \frac{1}{T}\sum_{t=1}^{T} h^{-1}K_\rho^J\left(\frac{x-X_t}{h}\right).$$

(b) Find the expression for β in terms of $\omega_K(\cdot,\rho)$ and α such that $\sup_{x\in[a,a+h)}|E\bar{g}(x)-g(x)| = O(h^2)$.

(c) 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 $x \in (b-h,b]$.