Lecture Notes on Nonparametrics

Bruce E. Hansen
University of Wisconsin

Spring 2009

1 Introduction

Parametric means finite-dimensional. **Non-parametric** means infinite-dimensional.

The differences are profound.

Typically, parametric estimates converge at a $n^{-1/2}$ rate. Non-parametric estimates typically converge at a rate slower than $n^{-1/2}$.

Typically, in parametric models there is no distinction between the true model and the fitted model. In contrast, non-parametric methods typically distinguish between the true and fitted models.

Non-parametric methods make the complexity of the fitted model depend upon the sample. The more information is in the sample (i.e., the larger the sample size), the greater the degree of complexity of the fitted model. Taking this seriously requires a distinct distribution theory.

Non-parametric theory acknowledges that fitted models are approximations, and therefore are inherently misspecified. Misspecification implies estimation bias. Typically, increasing the complexitiy of a fitted model decreases this bias but increases the estimation variance. Nonparametric methods acknowledge this trade-off and attempt to set model complexity to minimize an overall measure of fit, typically mean-squared error (MSE).

There are many nonparametric statistical objects of potential interest, including density functions (univariate and multivariate), density derivatives, conditional density functions, conditional distribution functions, regression functions, median functions, quantile functions, and variance functions. Sometimes these nonparametric objects are of direct interest. Sometimes they are of interest only as an input to a second-stage estimation problem. If this second-stage problem is described by a finite dimensional parameter we call the estimation problem **semiparametric**.

Nonparametric methods typically involve some sort of approximation or **smoothing** method. Some of the main methods are called **kernels**, **series**, and **splines**.

Nonparametric methods are typically indexed by a **bandwidth** or **tuning parameter** which controls the degree of complexity. The choice of bandwidth is often critical to implementation. Data-dependent rules for determination of the bandwidth are therefore essential for nonparametric methods. Nonparametric methods which require a bandwidth, but do not have an explicit data-dependent rule for selecting the bandwidth, are incomplete. Unfortunately this is quite common, due to the difficulty in developing rigorous rules for bandwidth selection. Often in these cases the bandwidth is selected based on a related statistical problem. This is a feasible yet worrisome compromise.

Many nonparametric problems are generalizations of univariate density estimation. We will start with this simple setting, and explore its theory in considerable detail.

2 Kernel Density Estimation

2.1 Discrete Estimator

Let X be a random variable with continuous distribution $F(x)$ and density $f(x) = \frac{d}{dx}F(x)$. The goal is to estimate $f(x)$ from a random sample $\{X_1, ..., X_n\}$.

The distribution function $F(x)$ is naturally estimated by the EDF $\hat{F}(x) = n^{-1}\sum_{i=1}^{n} 1\left(X_i \leq x\right).$ It might seem natural to estimate the density $f(x)$ as the derivative of $\hat{F}(x)$, $\frac{d}{dx}\hat{F}(x)$, but this estimator would be a set of mass points, not a density, and as such is not a useful estimate of $f(x)$.

Instead, consider a discrete derivative. For some small $h > 0$, let

$$\hat{f}(x) = \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h}$$

We can write this as

$$\begin{aligned} \frac{1}{2nh}\sum_{i=1}^{n} 1\left(x + h < X_i \leq x + h\right) &= \frac{1}{2nh}\sum_{i=1}^{n} 1\left(\frac{|X_i - x|}{h} \leq 1\right) \\ &= \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) \end{aligned}$$

where

$$k(u) = \begin{cases} \frac{1}{2}, & |u| \leq 1 \\ 0 & |u| > 1 \end{cases}$$

is the uniform density function on $[-1, 1]$.

The estimator $\hat{f}(x)$ counts the percentage of observations which are clsoe to the point x. If many observations are near x, then $\hat{f}(x)$ is large. Conversely, if only a few X_i are near x, then $\hat{f}(x)$ is small. The **bandwidth** h controls the degree of smoothing.

$\hat{f}(x)$ is a special case of what is called a kernel estimator. The general case is

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)$$

where $k(u)$ is a **kernel function**.

2.2 Kernel Functions

A **kernel function** $k(u) : \mathbb{R} \to \mathbb{R}$ is any function which satisfies $\int_{-\infty}^{\infty} k(u)du = 1$.

A **non-negative** kernel satisfies $k(u) \geq 0$ for all u. In this case, $k(u)$ is a probability density function.

The **moments** of a kernel are $\kappa_j(k) = \int_{-\infty}^{\infty} u^j k(u)du$.

A **symmetric** kernel function satisfies $k(u) = k(-u)$ for all u. In this case, all odd moments are zero. Most nonparametric estimation uses symmetric kernels, and we focus on this case.

The **order** of a kernel, ν, is defined as the order of the first non-zero moment. For example, if $\kappa_1(k) = 0$ and $\kappa_2(k) > 0$ then k is a second-order kernel and $\nu = 2$. If $\kappa_1(k) = \kappa_2(k) = \kappa_3(k) = 0$ but $\kappa_4(k) > 0$ then k is a fourth-order kernel and $\nu = 4$. The order of a symmetric kernel is always even.

Symmetric non-negative kernels are second-order kernels.

A kernel is **higher-order kernel** if $\nu > 2$. These kernels will have negative parts and are not probability densities. They are also refered to as **bias-reducing kernels**.

Common second-order kernels are listed in the following table

Table 1: Common Second-Order Kernels

Kernel	**Equation**	$R(k)$	$\kappa_2(k)$	$eff(k)$
Uniform	$k_0(u) = \frac{1}{2} 1 (\lvert u \rvert \leq 1)$	$1/2$	$1/3$	1.0758
Epanechnikov	$k_1(u) = \frac{3}{4} \left(1 - u^2\right) 1 (\lvert u \rvert \leq 1)$	$3/5$	$1/5$	1.0000
Biweight	$k_2(u) = \frac{15}{16} \left(1 - u^2\right)^2 1 (\lvert u \rvert \leq 1)$	$5/7$	$1/7$	1.0061
Triweight	$k_3(u) = \frac{35}{32} \left(1 - u^2\right)^3 1 (\lvert u \rvert \leq 1)$	$350/429$	$1/9$	1.0135
Gaussian	$k_\phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$	$1/2\sqrt{\pi}$	1	1.0513

In addition to the kernel formula we have listed its roughness $R(k)$, second moment $\kappa_2(k)$, and its efficiency $eff(k)$, the last which will be defined later. The **roughness** of a function is

$$R(g) = \int_{-\infty}^{\infty} g(u)^2 du.$$

The most commonly used kernels are the Epanechnikov and the Gaussian.

The kernels in the Table are special cases of the polynomial family

$$k_s(u) = \frac{(2s+1)!!}{2^{s+1} s!} \left(1 - u^2\right)^s 1 (\lvert u \rvert \leq 1)$$

where the double factorial means $(2s + 1)!! = (2s + 1)(2s - 1) \cdots 5 \cdot 3 \cdot 1$. The Gaussian kernel is obtained by taking the limit as $s \to \infty$ after rescaling. The kernels with higher s are smoother, yielding estimates $\hat{f}(x)$ which are smoother and possessing more derivatives. Estimates using the Gaussian kernel have derivatives of all orders.

For the purpose of nonparametric estimation the scale of the kernel is not uniquely defined. That is, for any kernel $k(u)$ we could have defined the alternative kernel $k^*(u) = b^{-1} k(u/b)$ for some constant $b > 0$. These two kernels are equivalent in the sense of producing the same density estimator, so long as the bandwidth is rescaled. That is, if $\hat{f}(x)$ is calculated with kernel k and bandwidth h, it is numerically identically to a calculation with kernel k^* and bandwidth $h^* = h/b$. Some authors use different definitions for the same kernels. This can cause confusion unless you are attentive.

Higher-order kernels are obtained by multiplying a second-order kernel by an $(\nu/2-1)$'th order polynomial in u^2. Explicit formulae for the general polynomial family can be found in B. Hansen (Econometric Theory, 2005), and for the Gaussian family in Wand and Schucany (Canadian Journal of Statistics, 1990). 4th and 6th order kernels of interest are given in Tables 2 and 3.

Table 2: Fourth-Order Kernels

Kernel	**Equation**	$R(k)$	$\kappa_4(k)$	$eff(k)$
Epanechnikov	$k_{4,1}(u) = \frac{15}{8}\left(1-\frac{7}{3}u^2\right)k_1(u)$	5/4	−1/21	1.0000
Biweight	$k_{4,2}(u) = \frac{7}{4}\left(1-3u^2\right)k_2(u)$	805/572	−1/33	1.0056
Triweight	$k_{4,3}(u) = \frac{27}{16}\left(1-\frac{11}{3}u^2\right)k_3(u)$	3780/2431	−3/143	1.0134
Gaussian	$k_{4,\phi}(u) = \frac{1}{2}\left(3-u^2\right)k_\phi(u)$	$27/32\sqrt{\pi}$	−3	1.0729

Table 3: Sixth-Order Kernels

Kernel	**Equation**	$R(k)$	$\kappa_6(k)$	$eff(k)$
Epanechnikov	$k_{6,1}(u) = \frac{175}{64}\left(1-6u^2+\frac{33}{5}u^4\right)k_1(u)$	1575/832	5/429	1.0000
Biweight	$k_{6,2}(u) = \frac{315}{128}\left(1-\frac{22}{3}u^2+\frac{143}{15}u^4\right)k_2(u)$	29295/14144	1/143	1.0048
Triweight	$k_{6,2}(u) = \frac{297}{128}\left(1-\frac{26}{3}u^2+13u^4\right)k_3(u)$	301455/134368	1/221	1.0122
Gaussian	$k_{6,\phi}(u) = \frac{1}{8}\left(15-10u^2+u^4\right)k_\phi(u)$	$2265/2048\sqrt{\pi}$	15	1.0871

2.3 Density Estimator

We now discuss some of the numerical properties of the kernel estimator

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)$$

viewed as a function of x.

First, if $k(u)$ is non-negative then it is easy to see that $\hat{f}(x) \geq 0$. However, this is not guarenteed if k is a higher-order kernel. That is, in this case it is possible that $\hat{f}(x) < 0$ for some values of x. When this happens it is prudent to zero-out the negative bits and then rescale:

$$\tilde{f}(x) = \frac{\hat{f}(x)1\left(\hat{f}(x) \geq 0\right)}{\int_{-\infty}^{\infty}\hat{f}(x)1\left(\hat{f}(x) \geq 0\right)dx}.$$

$\tilde{f}(x)$ is non-negative yet has the same asymptotic properties as $\hat{f}(x)$. Since the integral in the denominator is not analytically available this needs to be calculated numerically.

Second, $\hat{f}(x)$ integrates to one. To see this, first note that by the change-of-variables $u = (X_i - x)/h$ which has Jacobian h,

$$\int_{-\infty}^{\infty}\frac{1}{h}k\left(\frac{X_i - x}{h}\right)dx = \int_{-\infty}^{\infty}k\left(u\right)du = 1.$$

The change-of variables $u = (X_i - x)/h$ will be used frequently, so it is useful to be familiar with this transformation. Thus

$$\int_{-\infty}^{\infty} \hat{f}(x)dx = \int_{-\infty}^{\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}k\left(\frac{X_i - x}{h}\right) dx = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \frac{1}{h}k\left(\frac{X_i - x}{h}\right) dx = \frac{1}{n}\sum_{i=1}^{n} 1 = 1$$

as claimed. Thus $\hat{f}(x)$ is a valid density function when k is non-negative.

Third, we can also calculate the numerical moments of the density $\hat{f}(x)$. Again using the change-of-variables $u = (X_i - x)/h$, the mean of the estimated density is

$$\begin{aligned}
\int_{-\infty}^{\infty} x\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} x\frac{1}{h}k\left(\frac{X_i - x}{h}\right) dx \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} (X_i + uh)\, k(u)\, du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i \int_{-\infty}^{\infty} k(u)\, du + \frac{1}{n}\sum_{i=1}^{n} h \int_{-\infty}^{\infty} uk(u)\, du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i
\end{aligned}$$

the sample mean of the X_i.

The second moment of the estimated density is

$$\begin{aligned}
\int_{-\infty}^{\infty} x^2\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} x^2\frac{1}{h}k\left(\frac{X_i - x}{h}\right) dx \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} (X_i + uh)^2\, k(u)\, du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \frac{2}{n}\sum_{i=1}^{n} X_i h \int_{-\infty}^{\infty} k(u)du + \frac{1}{n}\sum_{i=1}^{n} h^2 \int_{-\infty}^{\infty} u^2k(u)\, du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\kappa_2(k).
\end{aligned}$$

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

$$\begin{aligned}
\int_{-\infty}^{\infty} x^2\hat{f}(x)dx - \left(\int_{-\infty}^{\infty} x\hat{f}(x)dx\right)^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\kappa_2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\
&= \hat{\sigma}^2 + h^2\kappa_2(k)
\end{aligned}$$

where $\hat{\sigma}^2$ is the sample variance. Thus the density estimate inflates the sample variance by the factor $h^2\kappa_2(k)$.

These are the numerical mean and variance of the estimated density $\hat{f}(x)$, not its sampling mean and variance.

2.4 Estimation Bias

It is useful to observe that expectations of kernel transformations can be written as integrals which take the form of a convolution of the kernel and the density function:

$$\mathrm{E}\frac{1}{h}k\left(\frac{X_i-x}{h}\right)=\int_{-\infty}^{\infty}\frac{1}{h}k\left(\frac{z-x}{h}\right)f(z)dz$$

Using the change-of variables $u=(z-x)/h$, this equals

$$\int_{-\infty}^{\infty}k\left(u\right)f(x+hu)du.$$

By the linearity of the estimator we see

$$\mathrm{E}\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\frac{1}{h}k\left(\frac{X_i-x}{h}\right)=\int_{-\infty}^{\infty}k\left(u\right)f(x+hu)du$$

The last expression shows that the expected value is an average of $f(z)$ locally about x.

This integral (typically) is not analytically solvable, so we approximate it using a Taylor expansion of $f(x+hu)$ in the argument hu, which is valid as $h\to 0$. For a ν'th-order kernel we take the expansion out to the ν'th term

$$\begin{aligned}f\left(x+hu\right) &= f(x)+f^{(1)}(x)hu+\frac{1}{2}f^{(2)}(x)h^2u^2+\frac{1}{3!}f^{(3)}(x)h^3u^3+\cdots\\ &\quad +\frac{1}{\nu!}f^{(\nu)}(x)h^{\nu}u^{\nu}+o\left(h^{\nu}\right).\end{aligned}$$

The remainder is of smaller order than h^{ν} as $h\to\infty$, which is written as $o(h^{\nu})$. (This expansion assumes $f^{(\nu+1)}(x)$ exists.)

Integrating term by term and using $\int_{-\infty}^{\infty}k\left(u\right)du=1$ and the definition $\int_{-\infty}^{\infty}k\left(u\right)u^{j}du=\kappa_j(k)$,

$$\begin{aligned}\int_{-\infty}^{\infty}k\left(u\right)f\left(x+hu\right)du &= f(x)+f^{(1)}(x)h\kappa_1(k)+\frac{1}{2}f^{(2)}(x)h^2\kappa_2(k)+\frac{1}{3!}f^{(3)}(x)h^3\kappa_3(k)+\cdots\\ &\quad +\frac{1}{\nu!}f^{(\nu)}(x)h^{\nu}\kappa_{\nu}(k)+o\left(h^{\nu}\right)\\ &= f(x)+\frac{1}{\nu!}f^{(\nu)}(x)h^{\nu}\kappa_{\nu}(k)+o\left(h^{\nu}\right)\end{aligned}$$

where the second equality uses the assumption that k is a ν'th order kernel (so $\kappa_j(k)=0$ for $j<\nu$).

This means that

$$\begin{aligned} \mathrm{E}\hat{f}(x) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\frac{1}{h}k\left(\frac{X_i - x}{h}\right) \\ &= f(x) + \frac{1}{\nu!}f^{(\nu)}(x)h^{\nu}\kappa_{\nu}(k) + o\left(h^{\nu}\right). \end{aligned}$$

The bias of $\hat{f}(x)$ is then

$$Bias(\hat{f}(x)) = \mathrm{E}\hat{f}(x) - f(x) = \frac{1}{\nu!}f^{(\nu)}(x)h^{\nu}\kappa_{\nu}(k) + o\left(h^{\nu}\right).$$

For second-order kernels, this simplifies to

$$Bias(\hat{f}(x)) = \frac{1}{2}f^{(2)}(x)h^{2}\kappa_{2}(k) + O\left(h^{4}\right).$$

For second-order kernels, the bias is increasing in the square of the bandwidth. Smaller bandwidths imply reduced bias. The bias is also proportional to the second derivative of the density $f^{(2)}(x)$. Intuitively, the estimator $\hat{f}(x)$ smooths data local to $X_i = x$, so is estimating a smoothed version of $f(x)$. The bias results from this smoothing, and is larger the greater the curvature in $f(x)$.

When higher-order kernels are used (and the density has enough derivatives), the bias is proportional to h^{ν}, which is of lower order than h^2. Thus the bias of estimates using higher-order kernels is of lower order than estimates from second-order kernels, and this is why they are called bias-reducing kernels. This is the advantage of higher-order kernels.

2.5 Estimation Variance

Since the kernel estimator is a linear estimator, and $k\left(\frac{X_i - x}{h}\right)$ is iid,

$$\begin{aligned} \operatorname{var}\left(\hat{f}(x)\right) &= \frac{1}{nh^2}\operatorname{var}\left(k\left(\frac{X_i - x}{h}\right)\right) \\ &= \frac{1}{nh^2}\mathrm{E}k\left(\frac{X_i - x}{h}\right)^2 - \frac{1}{n}\left(\frac{1}{h}\mathrm{E}k\left(\frac{X_i - x}{h}\right)\right)^2 \end{aligned}$$

From our analysis of bias we know that $\frac{1}{h}\mathrm{E}k\left(\frac{X_i - x}{h}\right) = f(x)+o(1)$ so the second term is $O\left(\frac{1}{n}\right)$. For the first term, write the expectation as an integral, make a change-of-variables and a first-order

Taylor expansion

$$\begin{aligned}
\frac{1}{h}\mathrm{E}k\left(\frac{X_i - x}{h}\right)^2 &= \frac{1}{h}\int_{-\infty}^{\infty} k\left(\frac{z-x}{h}\right)^2 f(z)dz \\
&= \int_{-\infty}^{\infty} k\left(u\right)^2 f\left(x+hu\right)du \\
&= \int_{-\infty}^{\infty} k\left(u\right)^2 \left(f\left(x\right) + O\left(h\right)\right)du \\
&= f\left(x\right)R(k) + O\left(h\right)
\end{aligned}$$

where $R(k) = \int_{-\infty}^{\infty} k\left(u\right)^2 du$ is the roughness of the kernel. Together, we see

$$\operatorname{var}\left(\hat{f}(x)\right) = \frac{f\left(x\right)R(k)}{nh} + O\left(\frac{1}{n}\right)$$

The remainder $O\left(\frac{1}{n}\right)$ is of smaller order than the $O\left(\frac{1}{nh}\right)$ leading term, since $h^{-1} \to \infty$.

2.6 Mean-Squared Error

A common and convenient measure of estimation precision is the mean-squared error

$$\begin{aligned}
MSE(\hat{f}(x)) &= \mathrm{E}\left(\hat{f}(x) - f(x)\right)^2 \\
&= Bias(\hat{f}(x))^2 + \operatorname{var}\left(\hat{f}(x)\right) \\
&\simeq \left(\frac{1}{\nu!}f^{(\nu)}(x)h^{\nu}\kappa_{\nu}(k)\right)^2 + \frac{f\left(x\right)R(k)}{nh} \\
&= \frac{\kappa_{\nu}^2(k)}{(\nu!)^2}f^{(\nu)}(x)^2 h^{2\nu} + \frac{f\left(x\right)R(k)}{nh} \\
&= AMSE(\hat{f}(x))
\end{aligned}$$

Since this approximation is based on asymptotic expansions this is called the asymptotic mean-squared-error (AMSE). Note that it is a function of the sample size n, the bandwidth h, the kernel function (through κ_ν and $R(k)$), and varies with x as $f^{(\nu)}(x)$ and $f(x)$ vary.

Notice as well that the first term (the squared bias) is increasing in h and the second term (the variance) is decreasing in nh. For $MSE(\hat{f}(x))$ to decline as $n \to \infty$ both of these terms must get small. Thus as $n \to \infty$ we must have $h \to 0$ and $nh \to \infty$. That is, the bandwidth must decrease, but not at a rate faster than sample size. This is sufficient to establish the pointwise consistency of the estimator. That is, for all x, $\hat{f}(x) \to_p f(x)$ as $n \to \infty$. We call this pointwise convergence as it is valid for each x individually. We discuss uniform convergence later.

A global measure of precision is the asymptotic mean integrated squared error (AMISE)

$$\begin{aligned} AMISE &= \int_{-\infty}^{\infty} AMSE(\hat{f}(x))dx \\ &= \frac{\kappa_\nu^2(k)}{(\nu!)^2} R\left(f^{(\nu)}\right) h^{2\nu} + \frac{R(k)}{nh}. \end{aligned}$$

where $R(f^{(\nu)}) = \int_{-\infty}^{\infty} \left(f^{(\nu)}(x)\right)^2 dx$ is the roughness of $f^{(\nu)}$.

2.7 Asymptotically Optimal Bandwidth

The AMISE formula expresses the MSE as a function of h. The value of h which minimizes this expression is called the asymptotically optimal bandwidth. The solution is found by taking the derivative of the AMISE with respect to h and setting it equal to zero:

$$\begin{aligned} \frac{d}{dh} AMISE &= \frac{d}{dh}\left(\frac{\kappa_\nu^2(k)}{(\nu!)^2} R\left(f^{(\nu)}\right) h^{2\nu} + \frac{R(k)}{nh}\right) \\ &= 2\nu h^{2\nu-1}\frac{\kappa_\nu^2(k)}{(\nu!)^2} R\left(f^{(\nu)}\right) - \frac{R(k)}{nh^2} \\ &= 0 \end{aligned}$$

with solution

$$\begin{aligned} h_0 &= C_\nu\left(k,f\right) n^{-1/(2\nu+1)} \\ C_\nu\left(k,f\right) &= R\left(f^{(\nu)}\right)^{-1/(2\nu+1)} A_\nu\left(k\right) \\ A_\nu\left(k\right) &= \left(\frac{(\nu!)^2 R(k)}{2\nu\kappa_\nu^2(k)}\right)^{1/(2\nu+1)} \end{aligned}$$

The optimal bandwidth is propotional to $n^{-1/(2\nu+1)}$. We say that the optimal bandwidth is of order $O\left(n^{-1/(2\nu+1)}\right)$. For second-order kernels the optimal rate is $O\left(n^{-1/5}\right)$. For higher-order kernels the rate is slower, suggesting that bandwidths are generally larger than for second-order kernels. The intuition is that since higher-order kernels have smaller bias, they can afford a larger bandwidth.

The constant of proportionality $C_\nu\left(k,f\right)$ depends on the kernel through the function $A_\nu\left(k\right)$ (which can be calculated from Table 1), and the density through $R(f^{(\nu)})$ (which is unknown).

If the bandwidth is set to h_0, then with some simplification the AMISE equals

$$AMISE_0\left(k\right) = (1+2\nu)\left(\frac{R\left(f^{(\nu)}\right)\kappa_\nu^2(k) R\left(k\right)^{2\nu}}{(\nu!)^2 \left(2\nu\right)^{2\nu}}\right)^{1/(2\nu+1)} n^{-2\nu/(2\nu+1)}.$$

For second-order kernels, this equals

$$AMISE_0\left(k\right) = \frac{5}{4}\left(\kappa_2^2(k)R(k)^4R\left(f^{(2)}\right)\right)^{1/5}n^{-4/5}.$$

As ν gets large, the convergence rate approaches the parametric rate n^{-1}. Thus, at least asymptotically, the slow convergence of nonparametric estimation can be mitigated through the use of higher-order kernels.

This seems a bit magical. What's the catch? For one, the improvement in convergence rate requires that the density is sufficiently smooth that derivatives exist up to the $(\nu+1)$'th order. As the density becomes increasingly smooth, it is easier to approximate by a low-dimensional curve, and gets closer to a parametric-type problem. This is exploiting the smoothness of f, which is inherently unknown. The other catch is that there is a some evidence that the benefits of higher-order kernels only develop when the sample size is fairly large. My sense is that in small samples, a second-order kernel would be the best choice, in moderate samples a 4th order kernel, and in larger samples a 6th order kernel could be used.

2.8 Asymptotically Optimal Kernel

Given that we have picked the kernel order, which kernel should we use? Examining the expression $AMISE_0$ we can see that for fixed ν the choice of kernel affects the asymptotic precision through the quantity $\kappa_\nu\left(k\right)R(k)^\nu$. All else equal, $AMISE$ will be minimized by selecting the kernel which minimizes this quantity. As we discussed earlier, only the shape of the kernel is important, not its scale, so we can set $\kappa_\nu = 1$. Then the problem reduces to minimization of $R(k) = \int_{-\infty}^{\infty} k(u)^2du$ subject to the constraints $\int_{-\infty}^{\infty} k(u)du = 1$ and $\int_{-\infty}^{\infty} u^\nu k(u)du = 1$. This is a problem in the calculus of variations. It turns out that the solution is a scaled of $k_{\nu,1}$ (see Muller (Annals of Statistics, 1984)). As the scale is irrelevant, this means that for estimation of the density function, the higher-order Epanechikov kernel $k_{\nu,1}$ with optimal bandwidth yields the lowest possible AMISE. For this reason, the Epanechikov kernel is often called the "optimal kernel".

To compare kernels, its relative efficiency is defined as

$$\begin{aligned} eff(k) &= \left(\frac{AMISE_0\left(k\right)}{AMISE_0\left(k_{\nu,1}\right)}\right)^{(1+2\nu)/2\nu} \\ &= \frac{\left(\kappa_\nu^2\left(k\right)\right)^{1/2\nu}R\left(k\right)}{\left(\kappa_\nu^2\left(k_{\nu,1}\right)\right)^{1/2\nu}R\left(k_{\nu,1}\right)} \end{aligned}$$

The ratios of the AMISE is raised to the power $\left(1+2\nu\right)/2\nu$ as for large n, the AMISE will be the same whether we use n observations with kernel $k_{\nu,1}$ or $n\ eff(k)$ observations with kernel k. Thus the penalty $eff(k)$ is expressed as a percentage of observations.

The efficiencies of the various kernels are given in Tables 1-3. Examining the second-order kernels, we see that relative to the Epanechnikov kernel, the uniform kernel pays a penalty of about 7%, the Gaussian kernel a penalty of about 5%, the Triweight kernel about 1.4%, and the Biweight kernel less than 1%. Examining the 4th and 6th-order kernels, we see that the relative efficiency of the Gaussian kernel deteriorates, while that of the Biweight and Triweight slightly improves.

The differences are not big. Still, the calculation suggests that the Epanechnikov and Biweight kernel classes are good choices for density estimation.

2.9 Rule-of-Thumb Bandwidth

The optimal bandwidth depends on the unknown quantity $R\left(f^{(\nu)}\right)$. Silverman proposed that we try the bandwidth computed by replacing $R\left(f^{(\nu)}\right)$ in the optimal formula by $R\left(g_{\hat{\sigma}}^{(\nu)}\right)$ where g_σ is a reference density – a plausible candidate for f, and $\hat{\sigma}^2$ is the sample standard deviation. The standard choice is to set $g_\sigma = \phi_{\hat{\sigma}}$, the $N(0, \hat{\sigma}^2)$ density. The idea is that if the true density is normal, then the computed bandwidth will be optimal. If the true density is reasonably close to the normal, then the bandwidth will be close to optimal. While not a perfect solution, it is a good place to start looking.

For any density g, if we set $g_\sigma(x) = \sigma^{-1}g(x/\sigma)$, then $g_\sigma^{(\nu)}(x) = \sigma^{-1-\nu}g^{(\nu)}(x/\sigma)$. Thus

$$\begin{aligned}
R\left(g_\sigma^{(\nu)}\right)^{-1/(2\nu+1)} &= \left(\int g_\sigma^{(\nu)}(x)^2 dx\right)^{-1/(2\nu+1)} \\
&= \left(\sigma^{-2-2\nu}\int g^{(\nu)}(x/\sigma)^2 dx\right)^{-1/(2\nu+1)} \\
&= \left(\sigma^{-1-2\nu}\int g^{(\nu)}(x)^2 dx\right)^{-1/(2\nu+1)} \\
&= \sigma R\left(g^{(\nu)}\right)^{-1/(2\nu+1)}.
\end{aligned}$$

Furthermore,

$$\left(R\left(\phi^{(\nu)}\right)\right)^{-1/(2\nu+1)} = 2\left(\frac{\pi^{1/2}\nu!}{(2\nu)!}\right)^{1/(2\nu+1)}.$$

Thus

$$R\left(\phi_{\hat{\sigma}}^{(\nu)}\right)^{-1/(2\nu+1)} = 2\hat{\sigma}\left(\frac{\pi^{1/2}\nu!}{(2\nu)!}\right)^{1/(2\nu+1)}.$$

The rule-of-thumb bandwidth is then $h = \hat{\sigma}C_\nu(k)\,n^{-1/(2\nu+1)}$ where

$$\begin{aligned}
C_\nu(k) &= R\left(\phi^{(\nu)}\right)^{-1/(2\nu+1)} A_\nu(k) \\
&= 2\left(\frac{\pi^{1/2}(\nu!)^3 R(k)}{2\nu\,(2\nu)!\kappa_\nu^2(k)}\right)^{1/(2\nu+1)}
\end{aligned}$$

We collect these constants in Table 4.

Table 4: Rule of Thumb Constants

Kernel	$\nu = 2$	$\nu = 4$	$\nu = 6$
Epanechnikov	2.34	3.03	3.53
Biweight	2.78	3.39	3.84
Triweight	3.15	3.72	4.13
Gaussian	1.06	1.08	1.08

Silverman Rule-of-Thumb: $h = \hat{\sigma} C_\nu(k)\, n^{-1/(2\nu+1)}$ where $\hat{\sigma}$ is the sample standard deviation, ν is the order of the kernel, and $C_\nu(k)$ is the constant from Table 4.

If a Gaussian kernel is used, this is often simplified to $h = \hat{\sigma} n^{-1/(2\nu+1)}$. In particular, for the standard second-order normal kernel, $h = \hat{\sigma} n^{-1/5}$.

2.10 Density Derivatives

Consider the problem of estimating the r'th derivative of the density:

$$f^{(r)}(x) = \frac{d^r}{dx^r} f(x).$$

A natural estimator is found by taking derivatives of the kernel density estimator. This takes the form

$$\hat{f}^{(r)}(x) = \frac{d^r}{dx^r} \hat{f}(x) = \frac{1}{nh^{1+r}} \sum_{i=1}^{n} k^{(r)}\left(\frac{X_i - x}{h}\right)$$

where

$$k^{(r)}(x) = \frac{d^r}{dx^r} k(x).$$

This estimator only makes sense if $k^{(r)}(x)$ exists and is non-zero. Since the Gaussian kernel has derivatives of all orders this is a common choice for derivative estimation.

The asymptotic analysis of this estimator is similar to that of the density, but with a couple of extra wrinkles and noticably different results. First, to calculate the bias we observe that

$$\mathrm{E}\frac{1}{h^{1+r}} k^{(r)}\left(\frac{X_i - x}{h}\right) = \int_{-\infty}^{\infty} \frac{1}{h^{1+r}} k^{(r)}\left(\frac{z - x}{h}\right) f(z) dz$$

To simplify this expression we use integration by parts. As the integral of $h^{-1} k^{(r)}\left(\frac{z-x}{h}\right)$ is $-k^{(r-1)}\left(\frac{z-x}{h}\right)$, we find that the above expression equals

$$\int_{-\infty}^{\infty} \frac{1}{h^r} k^{(r-1)}\left(\frac{z-x}{h}\right) f^{(1)}(z) dz.$$

Repeating this a total of r times, we obtain

$$\int_{-\infty}^{\infty} \frac{1}{h} k\left(\frac{z-x}{h}\right) f^{(r)}(z) dz.$$

Next, apply the change of variables to obtain

$$\int_{-\infty}^{\infty} k(u) f^{(r)}(x+hu)dz.$$

Now expand $f^{(r)}(x+hu)$ in a ν'th-order Taylor expansion about x, and integrate the terms to find that the above equals

$$f^{(r)}(x) + \frac{1}{\nu!} f^{(r+\nu)}(x) h^{\nu} \kappa_{\nu}(k) + o(h^{\nu})$$

where ν is the order of the kernel. Hence the asymptotic bias is

$$\begin{aligned} Bias(\hat{f}^{(r)}(x)) &= \mathrm{E}\hat{f}^{(r)}(x) - f^{(r)}(x) \\ &= \frac{1}{\nu!} f^{(r+\nu)}(x) h^{\nu} \kappa_{\nu}(k) + o(h^{\nu}). \end{aligned}$$

This of course presumes that f is differentiable of order at least $r + \nu + 1$.

For the variance, we find

$$\begin{aligned} \operatorname{var}\left(\hat{f}^{(r)}(x)\right) &= \frac{1}{nh^{2+2r}} \operatorname{var}\left(k^{(r)}\left(\frac{X_i - x}{h}\right)\right) \\ &= \frac{1}{nh^{2+2r}} \mathrm{E} k^{(r)}\left(\frac{X_i - x}{h}\right)^2 - \frac{1}{n}\left(\frac{1}{nh^{1+r}} \mathrm{E} k^{(r)}\left(\frac{X_i - x}{h}\right)\right)^2 \\ &= \frac{1}{nh^{2+2r}} \int_{-\infty}^{\infty} k^{(r)}\left(\frac{z - x}{h}\right)^2 f(z)dz - \frac{1}{n} f^{(r)}(x)^2 + O\left(\frac{1}{n}\right) \\ &= \frac{1}{nh^{1+2r}} \int_{-\infty}^{\infty} k^{(r)}(u)^2 f(x+hu)\, du + O\left(\frac{1}{n}\right) \\ &= \frac{f(x)}{nh^{1+2r}} \int_{-\infty}^{\infty} k^{(r)}(u)^2 du + O\left(\frac{1}{n}\right) \\ &= \frac{f(x) R(k^{(r)})}{nh^{1+2r}} + O\left(\frac{1}{n}\right). \end{aligned}$$

The AMSE and AMISE are

$$AMSE(\hat{f}^{(r)}(x)) = \frac{f^{(r+\nu)}(x)^2 h^{2\nu} \kappa_{\nu}^2(k)}{(\nu!)^2} + \frac{f(x) R(k^{(r)})}{nh^{1+2r}}$$

and

$$AMISE(\hat{f}^{(r)}(x)) = \frac{R\left(f^{(r+\nu)}\right) h^{2\nu} \kappa_{\nu}^2(k)}{(\nu!)^2} + \frac{R(k^{(r)})}{nh^{1+2r}}.$$

Note that the order of the bias is the same as for estimation of the density. But the variance is now of order $O\left(\frac{1}{nh^{1+2r}}\right)$ which is much larger than the $O\left(\frac{1}{nh}\right)$ found earlier.

The asymptotically optimal bandwidth is

$$\begin{aligned} h_r &= C_{r,\nu}\left(k,f\right) n^{-1/(1+2r+2\nu)} \\ C_{r,\nu}\left(k,f\right) &= R\left(f^{(r+\nu)}\right)^{-1/(1+2r+2\nu)} A_{r,\nu}\left(k\right) \\ A_{r,\nu}\left(k\right) &= \left(\frac{\left(1+2r\right)\left(\nu!\right)^2 R(k^{(r)})}{2\nu\kappa_\nu^2\left(k\right)}\right)^{1/(1+2r+2\nu)} \end{aligned}$$

Thus the optimal bandwidth converges at a slower rate than for density estimation. Given this bandwidth, the rate of convergence for the AMISE is $O\left(n^{-2\nu/(2r+2\nu+1)}\right)$, which is slower than the $O\left(n^{-2\nu/(2\nu+1)}\right)^{-4/5})$ rate when $r = 0$.

We see that we need a different bandwidth for estimation of derivatives than for estimation of the density. This is a common situation which arises in nonparametric analysis. The optimal amount of smoothing depends upon the object being estimated, and the goal of the analysis.

The AMISE with the optimal bandwidth is

$$AMISE(\hat{f}^{(r)}(x)) = (1+2r+2\nu)\left(\frac{\kappa_\nu^2\left(k\right)}{\left(\nu!\right)^2\left(1+2r\right)}\right)^{(2r+1)/(1+2r+2\nu)}\left(\frac{R\left(k^{(r)}\right)}{2\nu}\right)^{2\nu/(1+2r+2\nu)} n^{-2\nu/(1+2r+2\nu)}.$$

We can also ask the question of which kernel function is optimal, and this is addressed by Muller (1984). The problem amounts to minimizing $R\left(k^{(r)}\right)$ subject to a moment condition, and the solution is to set k equal to $k_{\nu,r+1}$, the polynomial kernel of ν'th order and exponent $r+1$. Thus to a first derivative it is optimal to use a member of the Biweight class and for a second derivative a member of the Triweight class.

The relative efficiency of a kernel k is then

$$\begin{aligned} eff(k) &= \left(\frac{AMISE_0\left(k\right)}{AMISE_0\left(k_{\nu,r+1}\right)}\right)^{(1+2\nu+2r)/2\nu} \\ &= \left(\frac{\kappa_\nu^2\left(k\right)}{\kappa_\nu^2\left(k_{\nu,r+1}\right)}\right)^{(1+2r)/2\nu}\frac{R\left(k^{(r)}\right)}{R\left(k_{\nu,r+1}^{(r)}\right)}. \end{aligned}$$

The relative efficiencies of the various kernels are presented in Table 5. (The Epanechnikov kernel is not considered as it is inappropriate for derivative estimation, and similarly the Biweight kernel for $r = 2$). In contrast to the case $r = 0$, we see that the Gaussian kernel is highly inefficient, with the efficiency loss increasing with r and ν. These calculations suggest that when estimating density derivatives it is important to use the appropriate kernel.

Table 5: Relative Efficiency $eff(k)$

		Biweight	Triweight	Gaussian
$r=1$	$\nu=2$	1.0000	1.0185	1.2191
	$\nu=4$	1.0000	1.0159	1.2753
	$\nu=6$	1.0000	1.0136	1.3156
$r=2$	$\nu=2$		1.0000	1.4689
	$\nu=4$		1.0000	1.5592
	$\nu=6$		1.0000	1.6275

The Silverman Rule-of-Thumb may also be applied to density derivative estimation. Again using the reference density $g_\sigma = \phi_\sigma$, we find the rule-of-thumb bandwidth is $h = C_{r,\nu}(k)\,\hat{\sigma} n^{-1/(2r+2\nu+1)}$ where

$$C_{r,\nu}(k) = 2\left(\frac{\pi^{1/2}(1+2r)(\nu!)^2(r+\nu)!R\left(k^{(r)}\right)}{2\nu\kappa_\nu^2(k)(2r+2\nu)!}\right)^{1/(2r+2\nu+1)}.$$

The constants $C_{r,v}$ are collected in Table 6. For all kernels, the constants $C_{r,\nu}$ are similar but slightly decreasing as r increases.

Table 6: Rule of Thumb Constants

		Biweight	Triweight	Gaussian
$r=1$	$\nu=2$	2.49	2.83	0.97
	$\nu=4$	3.18	3.49	1.03
	$\nu=6$	3.44	3.96	1.04
$r=2$	$\nu=2$		2.70	0.94
	$\nu=4$		3.35	1.00
	$\nu=6$		3.84	1.02

2.11 Multivariate Density Estimation

Now suppose that X_i is a q-vector and we want to estimate its density $f(x) = f(x_1, ..., x_q)$. A multivariate kernel estimator takes the form

$$\hat{f}(x) = \frac{1}{n|H|}\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right)$$

where $K(u)$ is a multivariate kernel function depending on a bandwidth vector $H = (h_1, ..., h_q)'$ and $|H| = h_1 h_2 \cdots h_q$. A multivariate kernel satisfies That is,

$$\int K(u)\,(du) = \int K(u) du_1 \cdots du_q = 1$$

Typically, $K(u)$ takes the product form:

$$K(u) = k(u_1)\,k(u_2)\cdots k(u_q)\,.$$

As in the univariate case, $\hat{f}(x)$ has the property that it integrates to one, and is non-negative if $K(u) \geq 0$. When $K(u)$ is a product kernel then the marginal densities of $\hat{f}(x)$ equal univariate kernel density estimators with kernel functions k and bandwidths h_j.

With some work, you can show that when $K(u)$ takes the product form, the bias of the estimator is

$$Bias(\hat{f}(x)) = \frac{\kappa_\nu(k)}{\nu!} \sum_{j=1}^{q} \frac{\partial^\nu}{\partial x_j^\nu} f(x) h_j^\nu + o\left(h_1^\nu + \cdots + h_q^\nu\right)$$

and the variance is

$$\begin{aligned} \operatorname{var}\left(\hat{f}(x)\right) &= \frac{f(x) R(K)}{n |H|} + O\left(\frac{1}{n}\right) \\ &= \frac{f(x) R(k)^q}{n h_1 h_2 \cdots h_q} + O\left(\frac{1}{n}\right). \end{aligned}$$

Hence the AMISE is

$$AMISE\left(\hat{f}(x)\right) = \frac{\kappa_\nu^2(k)}{(\nu!)^2} \int \left(\sum_{j=1}^{q} \frac{\partial^\nu}{\partial x_j^\nu} f(x) h_j^\nu \right)^2 (dx) + \frac{R(k)^q}{n h_1 h_2 \cdots h_q}$$

There is no closed-form solution for the bandwidth vector which minimizes this expression. However, even without doing do, we can make a couple of observations.

First, the AMISE depends on the kernel function only through $R(k)$ and $\kappa_\nu^2(k)$, so it is clear that for any given ν, the optimal kernel minimizes $R(k)$, which is the same as in the univariate case.

Second, the optimal bandwidths will all be of order $n^{-1/(2\nu+q)}$ and the optimal AMISE of order $n^{-2\nu/(2\nu+q)}$. This rates are slower than the univariate ($q = 1$) case. The fact that dimension has an adverse effect on convergence rates is called the **curse of dimensionality**. Many theoretical papers circumvent this problem through the following trick. Suppose you need the AMISE of the estimator to converge at a rate $O\left(n^{-1/2}\right)$ or faster. This requires $2\nu/\left(2\nu+q\right) > 1/2$, or $q < 2\nu$. For second-order kernels ($\nu = 2$) this restricts the dimension to be 3 or less. What some authors will do is slip in an assumption of the form: "Assume $f(x)$ is differentiable of order $\nu + 1$ where $\nu > q/2$," and then claim that their results hold for all q. The trouble is that what the author is doing is imposing greater smoothness as the dimension increases. This doesn't really avoid the curse of dimensionality, rather it hides it behind what appears to be a technical assumption. The bottom line is that nonparametric objects are much harder to estimate in higher dimensions, and that is why it is called a "curse".

To derive a rule-of-thumb, suppose that $h_1 = h_2 = \cdots = h_q = h$. Then

$$AMISE\left(\hat{f}(x)\right) = \frac{\kappa_\nu^2(k) R\left(\nabla^\nu f\right)}{(\nu!)^2} h^{2\nu} + \frac{R(k)^q}{n h^q}$$

where

$$\nabla^{\nu} f(x) = \sum_{j=1}^{q} \frac{\partial^{\nu}}{\partial x_j^{\nu}} f(x).$$

We find that the optimal bandwidth is

$$h_0 = \left(\frac{(\nu!)^2 \, qR(k)^q}{2\nu\kappa_\nu^2(k) R\left(\nabla^\nu f\right)} \right)^{1/(2\nu+q)} n^{-1/(2\nu+q)}$$

For a rule-of-thumb bandwidth, we replace f by the multivariate normal density ϕ. We can calculate that

$$R\left(\nabla^\nu \phi\right) = \frac{q}{\pi^{q/2} 2^{q+\nu}} \left((2\nu-1)!! + (q-1)\left((\nu-1)!!\right)^2 \right).$$

Making this substitution, we obtain $h_0 = C_\nu\left(k,q\right) n^{-1/(2\nu+q)}$where

$$C_\nu\left(k,q\right) = \left(\frac{\pi^{q/2} 2^{q+\nu-1} \left(\nu!\right)^2 R(k)^q}{\nu\kappa_\nu^2(k)\left((2\nu-1)!! + (q-1)\left((\nu-1)!!\right)^2 \right)} \right)^{1/(2\nu+q)}.$$

Now this assumed that all variables had unit variance. Rescaling the bandwidths by the standard deviation of each variable, we obtain the rule-of-thumb bandwidth for the j'th variable:

$$h_j = \hat{\sigma}_j C_\nu\left(k,q\right) n^{-1/(2\nu+q)}.$$

Numerical values for the constants $C_\nu(k,q)$ are given in Table 7 for $q = 2, 3, 4$.

Table 7: Rule of Thumb Constants

$\nu = 2$	$q = 2$	$q = 3$	$q = 4$
Epanechnikov	2.20	2.12	2.07
Biweight	2.61	2.52	2.46
Triweight	2.96	2.86	2.80
Gaussian	1.00	0.97	0.95
$\nu = 4$			
Epanechnikov	3.12	3.20	3.27
Biweight	3.50	3.59	3.67
Triweight	3.84	3.94	4.03
Gaussian	1.12	1.16	1.19
$\nu = 6$			
Epanechnikov	3.69	3.83	3.96
Biweight	4.02	4.18	4.32
Triweight	4.33	4.50	4.66
Gaussian	1.13	1.18	1.23

2.12 Least-Squares Cross-Validation

Rule-of-thumb bandwidths are a useful starting point, but they are inflexible and can be far from optimal.

Plug-in methods take the formula for the optimal bandwidth, and replace the unknowns by estimates, e.g. $R\left(\hat{f}^{(\nu)}\right)$. But these initial estimates themselves depend on bandwidths. And each situation needs to be individually studied. Plug-in methods have been thoroughly studied for univariate density estimation, but are less well developed for multivariate density estimation and other contexts.

A flexible and generally applicable data-dependent method is cross-validation. This method attempts to make a direct estimate of the squared error, and pick the bandwidth which minimizes this estimate. In many senses the idea is quite close to model selection based on a information criteria, such as Mallows or AIC.

Given a bandwidth h and density estimate $\hat{f}(x)$ of $f(x)$, define the mean integrated squared error (MISE)

$$MISE\,(h) = \int \left(\hat{f}(x) - f(x)\right)^2 (dx) = \int \hat{f}(x)^2 \,(dx) - 2\int \hat{f}(x) f(x) \,(dx) + \int f(x)^2 \,(dx)$$

Optimally, we want $\hat{f}(x)$ to be as close to $f(x)$ as possible, and thus for $MISE\,(h)$ to be as small as possible.

As $MISE\,(h)$ is unknown, cross-validation replaces it with an estimate.

The goal is to find an estimate of $MISE\,(h)$, and find the h which minimizes this estimate.

As the third term in the above expression does not depend on the bandwidth h, it can be ignored.

The first term can be directly calculated.

For the univariate case

$$\begin{aligned}\int \hat{f}(x)^2 dx &= \int \left(\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)\right)^2 dx \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int k\left(\frac{X_i - x}{h}\right) k\left(\frac{X_j - x}{h}\right) dx\end{aligned}$$

The convolution of k with itself is $\bar{k}(x) = \int k\,(u)\,k\,(x-u)\,du = \int k\,(u)\,k\,(u-x)\,du$ (by symmetry of k). Then making the change of variables $u = \dfrac{X_i - x}{h}$,

$$\begin{aligned}\frac{1}{h}\int k\left(\frac{X_i - x}{h}\right) k\left(\frac{X_i - x}{h}\right) dx &= \int k\,(u)\,k\left(u - \frac{X_i - X_j}{h}\right) du \\ &= \bar{k}\left(\frac{X_i - X_j}{h}\right).\end{aligned}$$

Hence

$$\int \hat{f}(x)^2 dx = \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} \bar{k}\left(\frac{X_i - X_j}{h}\right).$$

Discussion of $\bar{k}\left(x\right)$ can be found in the following section.

In the multivariate case,

$$\int \hat{f}(x)^2 dx = \frac{1}{n^2 \left|H\right|} \sum_{i=1}^{n} \sum_{j=1}^{n} \bar{K}\left(H^{-1}\left(X_i - X_j\right)\right)$$

where $\bar{K}\left(u\right) = \bar{k}\left(u_1\right) \cdots \bar{k}\left(u_q\right)$

The second term in the expression for $MISE\left(h\right)$ depends on $f(x)$ so is unknown and must be estimated. An integral with respect to $f(x)$ is an expectation with respect to the random variable X_i. While we don't know the true expectation, we have the sample, so can estimate this expectation by taking the sample average. In general, a reasonable estimate of the integral $\int g(x)f(x)dx$ is $\frac{1}{n}\sum_{i=1}^{n} g\left(X_i\right)$, suggesting the estimate $\frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(X_i\right)$. In this case, however, the function $\hat{f}\left(x\right)$ is itself a function of the data. In particular, it is a function of the observation X_i. A way to clean this up is to replace $\hat{f}\left(X_i\right)$ with the "leave-one-out" estimate $\hat{f}_{-i}\left(X_i\right)$, where

$$\hat{f}_{-i}\left(x\right) = \frac{1}{\left(n-1\right)\left|H\right|} \sum_{j \neq i} K\left(H^{-1}\left(X_j - x\right)\right)$$

is the density estimate computed without observation X_i, and thus

$$\hat{f}_{-i}\left(X_i\right) = \frac{1}{\left(n-1\right)\left|H\right|} \sum_{j \neq i} K\left(H^{-1}\left(X_j - X_i\right)\right).$$

That is, $\hat{f}_{-i}\left(X_i\right)$ is the density estimate at $x = X_i$, computed with the observations except X_i. We end up suggesting to estimate $\int \hat{f}(x)f(x)dx$ with

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}\left(X_i\right) = \frac{1}{n\left(n-1\right)\left|H\right|} \sum_{i=1}^{n} \sum_{j \neq i} K\left(H^{-1}\left(X_j - X_i\right)\right)$$

. It turns out that this is an unbiased estimate, in the sense that

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}\left(X_i\right)\right) = \mathrm{E}\left(\int \hat{f}(x)f(x)dx\right)$$

To see this, the LHS is

$$\begin{aligned}
\mathrm{E}\hat{f}_{-n}\left(X_n\right) &= \mathrm{E}\left(\mathrm{E}\left(\hat{f}_{-n}\left(X_n\right) \mid X_1, ..., X_{n-1}\right)\right) \\
&= \mathrm{E}\left(\int \hat{f}_{-n}(x) f(x)\,(dx)\right) \\
&= \int \mathrm{E}\left(\hat{f}(x)\right) f(x)\,(dx) \\
&= \mathrm{E}\left(\int \hat{f}(x) f(x)\,(dx)\right)
\end{aligned}$$

the second-to-last equality exchanging integration, and since $\mathrm{E}\left(\hat{f}(x)\right)$ depends only in the bandwidth, not the sample size.

Together, the least-squares cross-validation criterion is

$$CV\left(h_1, ..., h_q\right) = \frac{1}{n^2 |H|} \sum_{i=1}^{n} \sum_{j=1}^{n} \bar{K}\left(H^{-1}\left(X_i - X_j\right)\right) - \frac{2}{n(n-1)|H|} \sum_{i=1}^{n} \sum_{j \neq i} K\left(H^{-1}\left(X_j - X_i\right)\right).$$

Another way to write this is

$$\begin{aligned}
CV\left(h_1, ..., h_q\right) &= \frac{\bar{K}(0)}{n|H|} + \frac{1}{n^2|H|} \sum_{i=1}^{n} \sum_{j \neq i} \bar{K}\left(H^{-1}\left(X_i - X_j\right)\right) - \frac{2}{n(n-1)|H|} \sum_{i=1}^{n} \sum_{j \neq i} K\left(H^{-1}\left(X_j - X_i\right)\right) \\
&\simeq \frac{R(k)^q}{n|H|} + \frac{1}{n^2|H|} \sum_{i=1}^{n} \sum_{j \neq i} \left(\bar{K}\left(H^{-1}\left(X_i - X_j\right)\right) - 2K\left(H^{-1}\left(X_j - X_i\right)\right)\right)
\end{aligned}$$

using $\bar{K}(0) = \bar{k}(0)^q$ and $\bar{k}(0) = \int k(u)^2$, and the approximation is by replacing $n-1$ by n.

The cross-validation bandwidth vector are the value $\hat{h}_1, ..., \hat{h}_q$ which minimizes $CV\left(h_1, ..., h_q\right)$. The cross-validation function is a complicated function of the bandwidhts, so this needs to be done numerically.

In the univariate case, h is one-dimensional this is typically done by plotting (a grid search). Pick a lower and upper value $[h_1, h_2]$, define a grid on this set, and compute $CV(h)$ for each h in the grid. A plot of $CV(h)$ against h is a useful diagnostic tool.

The $CV(h)$ function can be misleading for small values of h. This arises when there is data rounding. Some authors define the cross-validation bandwidth as the largest local minimer of $CV(h)$ (rather than the global minimizer). This can also be avoided by picking a sensible initial range $[h_1, h_2]$. The rule-of-thumb bandwidth can be useful here. If h_0 is the rule-of-thumb bandwidth, then use $h_1 = h_0/3$ and $h_2 = 3h_0$ or similar.

We we discussed above, $CV\left(h_1, ..., h_q\right) + \int f(x)^2\,(dx)$ is an unbiased estimate of $MISE(h)$. This by itself does not mean that $\hat{h}$ is a good estimate of h_0, the minimizer of $MISE(h)$, but it turns out that this is indeed the case. That is,

$$\frac{\hat{h}-h_0}{h_0} \rightarrow_p 0$$

Thus, $\hat{h}$ is asymptotically close to h_0, but the rate of convergence is very slow.

The CV method is quite flexible, as it can be applied for any kernel function.

If the goal, however, is estimation of density derivatives, then the CV bandwidth $\hat{h}$ is not appropriate. A practical solution is the following. Recall that the asymptotically optimal bandwidth for estimation of the density takes the form $h_0 = C_\nu(k,f)\, n^{-1/(2\nu+1)}$ and that for the r'th derivative is $h_r = C_{r,\nu}(k,f)\, n^{-1/(1+2r+2\nu)}$. Thus if the CV bandwidth $\hat{h}$ is an estimate of h_0, we can estimate $C_\nu(k,f)$ by $\hat{C}_\nu = \hat{h}n^{1/(2\nu+1)}$. We also saw (at least for the normal reference family) that $C_{r,\nu}(k,f)$ was relatively constant across r. Thus we can replace $C_{r,\nu}(k,f)$ with $\hat{C}_\nu$ to find

$$\begin{aligned}
\hat{h}_r &= \hat{C}_\nu n^{-1/(1+2r+2\nu)} \\
&= \hat{h}n^{1/(2\nu+1)-1/(1+2r+2\nu)} \\
&= \hat{h}n^{(1+2r+2\nu)/(2\nu+1)(1+2r+2\nu)-(2\nu+1)/(1+2r+2\nu)(2\nu+1)} \\
&= \hat{h}n^{2r/((2\nu+1)(1+2r+2\nu))}
\end{aligned}$$

Alternatively, some authors use the rescaling

$$\hat{h}_r = \hat{h}^{(1+2\nu)/(1+2r+2\nu)}$$

2.13 Convolution Kernels

If $k(x) = \phi(x)$ then $\bar{k}(x) = \exp(-x^2/4)/\sqrt{4\pi}$.

When $k(x)$ is a higher-order Gaussian kernel, Wand and Schucany (Canadian Journal of Statistics, 1990, p. 201) give an expression for $\bar{k}(x)$.

For the polynomial class, because the kernel $k(u)$ has support on $[-1,1]$, it follows that $\bar{k}(x)$ has support on $[-2,2]$ and for $x \geq 0$ equals $\bar{k}(x) = \int_{x-1}^{1} k(u)k(x-u)du$. This integral can be easily solved using algebraic software (Maple, Mathematica), but the expression can be rather cumbersome.

For the 2nd order Epanechnikov, Biweight and Triweight kernels, for $0 \leq x \leq 2$,

$$\bar{k}_1(x) = \frac{3}{160}(2-x)^3\left(x^2+6x+4\right)$$

$$\bar{k}_2(x) = \frac{5}{3584}(2-x)^5\left(x^4+10x^3+36x^2+40x+16\right)$$

$$\bar{k}_3(x) = \frac{35}{1757\,184}(2-x)^7\left(5x^6+70x^5+404x^4+1176x^3+1616x^2+1120x+320\right)$$

These functions are symmetric, so the values for $x < 0$ are found by $\bar{k}(x) = \bar{k}(-x)$.

For the 4th, and 6th order Epanechnikov kernels, for $0 \leq x \leq 2$,

$$\bar{k}_{4,1}(x) = \frac{5}{2048}(2-x)^3\left(7x^6+42x^5+48x^4-160x^3-144x^2+96x+64\right)$$

$$\bar{k}_{6,1}(x) = \frac{105}{3407\,872}(2-x)^3\left(495x^{10}+2970x^9+2052x^8-19\,368x^7-32\,624x^6\right.$$
$$\left.+53\,088x^5+68\,352x^4-48\,640x^3-46\,720x^2+11\,520x+7680\right)$$

2.14 Asymptotic Normality

The kernel estimator is the sample average

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right).$$

We can therefore apply the central limit theorem.

But the convergence rate is not $\sqrt{n}$. We know that

$$\operatorname{var}\left(\hat{f}(x)\right) = \frac{f(x)\,R(k)^q}{nh_1h_2\cdots h_q}+O\left(\frac{1}{n}\right).$$

so the convergence rate is $\sqrt{nh_1h_2\cdots h_q}$. When we apply the CLT we scale by this, rather than the conventional $\sqrt{n}$.

As the estimator is biased, we also center at its expectation, rather than the true value

Thus

$$\begin{aligned}\sqrt{nh_1h_2\cdots h_q}\left(\hat{f}(x)-E\hat{f}(x)\right) &= \frac{\sqrt{nh_1h_2\cdots h_q}}{n}\sum_{i=1}^{n}\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right)-E\left(\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right)\right)\\ &= \frac{\sqrt{h_1h_2\cdots h_q}}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right)-E\left(\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right)\right)\right)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_{ni}\end{aligned}$$

where

$$Z_{ni} = \sqrt{h_1h_2\cdots h_q}\left(\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right)-E\left(\frac{1}{|H|}K\left(H^{-1}(X_i-x)\right)\right)\right)$$

We see that

$$\operatorname{var}(Z_{ni}) \simeq f(x)\,R(k)^q$$

Hence by the CLT,

$$\sqrt{nh_1h_2\cdots h_q}\left(\hat{f}(x)-E\hat{f}(x)\right)\rightarrow_d N\left(0,f(x)\,R(k)^q\right).$$

We also know that

$$E(\hat{f}(x)) = f(x) + \frac{\kappa_\nu(k)}{\nu!}\sum_{j=1}^{q}\frac{\partial^\nu}{\partial x_j^\nu}f(x)h_j^\nu + o\left(h_1^\nu + \cdots + h_q^\nu\right)$$

So another way of writing this is

$$\sqrt{nh_1h_2\cdots h_q}\left(\hat{f}(x) - f(x) - \frac{\kappa_\nu(k)}{\nu!}\sum_{j=1}^{q}\frac{\partial^\nu}{\partial x_j^\nu}f(x)h_j^\nu\right) \to_d N\left(0, f\left(x\right)R(k)^q\right).$$

In the univariate case this is

$$\sqrt{nh}\left(\hat{f}(x) - f(x) - \frac{\kappa_\nu(k)}{\nu!}f^{(2)}(x)h^\nu\right) \to_d N\left(0, f\left(x\right)R(k)\right)$$

This expression is most useful when the bandwidth is selected to be of optimal order, that is $h = Cn^{-1/(2\nu+1)}$, for then $\sqrt{nh}h^\nu = C^{\nu+1/2}$ and we have the equivalent statement

$$\sqrt{nh}\left(\hat{f}(x) - f(x)\right) \to_d N\left(C^{\nu+1/2}\frac{\kappa_\nu(k)}{\nu!}f^{(2)}(x), f\left(x\right)R(k)\right)$$

This says that the density estimator is asymptotically normal, with a non-zero asymptotic bias and variance.

Some authors play a dirty trick, by using the assumption that h is of smaller order than the optimal rate, e.g. $h = o\left(n^{-1/(2\nu+1)}\right)$. For then then obtain the result

$$\sqrt{nh}\left(\hat{f}(x) - f(x)\right) \to_d N\left(0, f\left(x\right)R(k)\right)$$

This appears much nicer. The estimator is asymptotically normal, with mean zero! There are several costs. One, if the bandwidth is really seleted to be sub-optimal, the estimator is simply less precise. A sub-optimal bandwidth results in a slower convergence rate. This is not a good thing. The reduction in bias is obtained at in increase in variance. Another cost is that the asymptotic distribution is misleading. It suggests that the estimator is unbiased, which is not honest. Finally, it is unclear how to pick this sub-optimal bandwidth. I call this assumption a dirty trick, because it is slipped in by authors to make their results cleaner and derivations easier. This type of assumption should be avoided.

2.15 Pointwise Confidence Intervals

The asymptotic distribution may be used to construct pointwise confidence intervals for $f(x)$. In the univariate case conventional confidence intervals take the form

$$\hat{f}(x) \pm 2\left(\hat{f}\left(x\right)R(k)/\left(nh\right)\right)^{1/2}.$$

These are not necessarily the best choice, since the variance equals the mean. This set has the unfortunate property that it can contain negative values, for example.

Instead, consider constructing the confidence interval by inverting a test statistic. To test $H_0 : f(x) = f_0$, a t-ratio is

$$t\left(f_0\right) = \frac{\hat{f}(x) - f_0}{\sqrt{nhf_0R(k)}}.$$

We reject H_0 if $\left|t\left(f_0\right)\right| > 2$. By the no-rejection rule, an asymptotic 95% confidence interval for f is the set of f_0 which do reject, i.e. the set of f such that $\left|t\left(f\right)\right| \leq 2$. This is

$$C(x) = \left\{ f : \left| \frac{\hat{f}(x) - f}{\sqrt{nhfR(k)}} \right| \leq 2 \right\}$$

This set must be found numerically.

3 Nonparametric Regression

3.1 Nadaraya-Watson Regression

Let the data be (y_i, X_i) where y_i is real-valued and X_i is a q-vector, and assume that all are continuously distributed with a joint density $f(y,x)$. Let $f(y \mid x) = f(y,x)/f(x)$ be the conditional density of y_i given X_i where $f(x) = \int f(y,x)\,dy$ is the marginal density of X_i. The regression function for y_i on X_i is

$$g(x) = E(y_i \mid X_i = x).$$

We want to estimate this nonparametrically, with minimal assumptions about g.

If we had a large number of observations where X_i exactly equals x, we could take the average value of the $y_i's$ for these observations. But since X_i is continously distributed, we won't observe multiple observations equalling the same value.

The solution is to consider a neighborhood of x, and note that if X_i has a positive density at x, we should observe a number of observations in this neighborhood, and this number is increasing with the sample size. If the regression function $g(x)$ is continuous, it should be reasonable constant over this neighborhood (if it is small enough), so we can take the average the the y_i values for these observations. The obvious trick is is to determine the size of the neighborhood to trade off the variation in $g(x)$ over the neighborhood (estimation bias) against the number of observations in the neighborhood (estimation variance).

we will observe a large number of X_i in any given neighborhood of x_i.

Take the one-regressor case $q = 1$.

Let a neighborhood of x be $x \pm h$ for some bandwidth $h > 0$. Then a simple nonparametric estimator of $g(x)$ is the average value of the $y_i's$ for the observations i such that X_i is in this neighborhood, that is,

$$\begin{aligned}\hat{g}(x) &= \frac{\sum_{i=1}^n 1\left(|X_i - x| \le h\right) y_i}{\sum_{i=1}^n 1\left(|X_i - x| \le h\right)} \\ &= \frac{\sum_{i=1}^n k\left(\dfrac{X_i - x}{h}\right) y_i}{\sum_{i=1}^n k\left(\dfrac{X_i - x}{h}\right)}\end{aligned}$$

where $k(u)$ is the uniform kernel.

In general, the kernel regression estimator takes this form, where $k(u)$ is a kernel function. It is known as the Nadaraya-Watson estimator, or local constant estimator.

When $q > 1$ the estimator is

$$\hat{g}(x) = \frac{\sum_{i=1}^n K\left(H^{-1}(X_i - x)\right) y_i}{\sum_{i=1}^n K\left(H^{-1}(X_i - x)\right)}$$

where $K(u)$ is a multivariate kernel function.

As an alternative motivation, note that the regression function can be written as

$$g(x) = \frac{\int y f(y,x)\,dy}{f(x)}$$

where $f(x) = \int f(y,x)\,dy$ is the marginal density of X_i. Now consider estimating g by replacing the density functions by the nonparametric estimates we have already studied. That is,

$$\hat{f}(y,x) = \frac{1}{n\,|H|\,h_y} \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) k\left(\frac{y_i - y}{h_y}\right)$$

where h_y is a bandwidth for smoothing in the y-direction. Then

$$\begin{aligned}
\hat{f}(x) &= \int \hat{f}(y,x)\,dy \\
&= \frac{1}{n\,|H|\,h_y} \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) \int k\left(\frac{y_i - y}{h_y}\right) dy \\
&= \frac{1}{n\,|H|} \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right)
\end{aligned}$$

and

$$\begin{aligned}
\int y\hat{f}(y,x)\,dy &= \frac{1}{n\,|H|\,h_y} \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) \int yk\left(\frac{y_i - y}{h_y}\right) dy \\
&= \frac{1}{n\,|H|} \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) y_i
\end{aligned}$$

and thus taking the ratio

$$\begin{aligned}
\hat{g}(x) &= \frac{\frac{1}{n\,|H|} \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) y_i}{\frac{1}{n\,|H|} \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right)} \\
&= \frac{\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) y_i}{\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right)}
\end{aligned}$$

again obtaining the Nadaraya-Watson estimator. Note that the bandwidth h_y has disappeared.

The estimator is ill-defined for values of x such that $\hat{f}(x) \leq 0$. This can occur in the tails of the distribution of X_i. As higher-order kernels can yield $\hat{f}(x) < 0$, many authors suggest using only second-order kernels for regression. I am unsure if this is a correct recommendation. If a higher-order kernel is used and for some x we find $\hat{f}(x) < 0$, this suggests that the data is so sparse in that neighborhood of x that it is unreasonable to estimate the regression function. It does not require the abandonment of higher-order kernels. We will follow convention and typically assume that k is second order ($\nu = 2$) for our presentation.

3.2 Asymptotic Distribution

We analyze the asymptotic distribution of the NW estimator $\hat{g}(x)$ for the case $q = 1$.

Since $E(y_i \mid X_i) = g(X_i)$, we can write the regression equation as $y_i = g(X_i) + e_i$ where $E(e_i \mid X_i) = 0$. We can also write the conditional variance as $E\left(e_i^2 \mid X_i = x\right) = \sigma^2(x)$.

Fix x. Note that

$$\begin{aligned} y_i &= g(X_i) + e_i \\ &= g(x) + (g(X_i) - g(x)) + e_i \end{aligned}$$

and therefore

$$\begin{aligned} \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) y_i &= \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) g(x) \\ &\quad + \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) (g(X_i) - g(x)) \\ &\quad + \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) e_i \\ &= \hat{f}(x)g(x) + \hat{m}_1(x) + \hat{m}_2(x), \end{aligned}$$

say. It follows that

$$\hat{g}(x) = g(x) + \frac{\hat{m}_1(x)}{\hat{f}(x)} + \frac{\hat{m}_2(x)}{\hat{f}(x)}.$$

We now analyze the asymptotic distributions of the components $\hat{m}_1(x)$ and $\hat{m}_2(x)$.

First take $\hat{m}_2(x)$. Since $E(e_i \mid X_i) = 0$ it follows that $E\left(k\left(\frac{X_i - x}{h}\right) e_i\right) = 0$ and thus $E(\hat{m}_2(x)) = 0$. Its variance is

$$\begin{aligned} var\left(\hat{m}_2(x)\right) &= \frac{1}{nh^2} E\left(k\left(\frac{X_i - x}{h}\right) e_i\right)^2 \\ &= \frac{1}{nh^2} E\left(k\left(\frac{X_i - x}{h}\right)^2 \sigma^2(X_i)\right) \end{aligned}$$

(by conditioning), and this is

$$\frac{1}{nh^2}\int k\left(\frac{z - x}{h}\right)^2 \sigma^2(z) f(z) dz$$

(where $f(z)$ is the density of X_i). Making the change of variables, this equals

$$\begin{aligned}\frac{1}{nh}\int k\left(u\right)^2\sigma^2(x+hu)f(x+hu)du &= \frac{1}{nh}\int k\left(u\right)^2\sigma^2(x)f(x)du+o\left(\frac{1}{nh}\right)\\ &= \frac{R(k)\sigma^2(x)f(x)}{nh}+o\left(\frac{1}{nh}\right)\end{aligned}$$

if $\sigma^2(x)f(x)$ are smooth in x. We can even apply the CLT to obtain that as $h\to 0$ and $nh\to\infty$,

$$\sqrt{nh}\hat{m}_2(x)\to_d N\left(0,R(k)\sigma^2(x)f(x)\right).$$

Now take $\hat{m}_1(x)$. Its mean is

$$\begin{aligned}E\hat{m}_1(x) &= \frac{1}{h}Ek\left(\frac{X_i-x}{h}\right)(g(X_i)-g(x))\\ &= \frac{1}{h}\int k\left(\frac{z-x}{h}\right)(g(z)-g(x))\,f(z)dz\\ &= \int k\left(u\right)(g(x+hu)-g(x))\,f(x+hu)du\end{aligned}$$

Now expanding both g and f in Taylor expansions, this equals, up to $o(h^2)$

$$\begin{aligned}&\int k\left(u\right)\left(uhg^{(1)}(x)+\frac{u^2h^2}{2}g^{(2)}(x)\right)\left(f(x)+uhf^{(1)}(x)\right)du\\ =\ &\left(\int k\left(u\right)udu\right)hg^{(1)}(x)f(x)\\ &+\left(\int k\left(u\right)u^2du\right)h^2\left(\frac{1}{2}g^{(2)}(x)f(x)+g^{(1)}(x)f^{(1)}(x)\right)\\ =\ &h^2\kappa_2\left(\frac{1}{2}g^{(2)}(x)f(x)+g^{(1)}(x)f^{(1)}(x)\right)\\ =\ &h^2\kappa_2B(x)f(x),\end{aligned}$$

where

$$B(x)=\frac{1}{2}g^{(2)}(x)+f(x)^{-1}g^{(1)}(x)f^{(1)}(x)$$

(If k is a higher-order kernel, this is $O(h^\nu)$ instead.) A similar expansion shows that $var(\hat{m}_1(x))=O\left(\frac{h^2}{nh}\right)$ which is of smaller order than $O\left(\frac{1}{nh}\right)$. Thus

$$\sqrt{nh}\left(\hat{m}_1(x)-h^2\kappa_2B(x)f(x)\right)\to_p 0$$

and since $\hat{f}(x)\to_p f(x)$,

$$\sqrt{nh}\left(\frac{\hat{m}_1(x)}{\hat{f}(x)}-h^2\kappa_2B(x)\right)\to_p 0$$

In summary, we have

$$\begin{aligned}
\sqrt{nh}\left(\hat{g}(x)-g(x)-h^2\kappa_2 B(x)\right) &= \sqrt{nh}\left(\frac{\hat{m}_1(x)}{\hat{f}(x)}-h^2\kappa_2 B(x)\right)+\frac{\sqrt{nh}\hat{m}_2(x)}{\hat{f}(x)} \\
&\xrightarrow{d} \frac{N\left(0,R(k)\sigma^2(x)f(x)\right)}{f(x)} \\
&= N\left(0,\frac{R(k)\sigma^2(x)}{f(x)}\right)
\end{aligned}$$

When X_i is a q-vector, the result is

$$\sqrt{n\left|H\right|}\left(\hat{g}(x)-g(x)-\kappa_2\sum_{j=1}^{q}h_j^2 B_j(x)\right)\xrightarrow{d}= N\left(0,\frac{R(k)^q\sigma^2(x)}{f(x)}\right)$$

where

$$B_j(x)=\frac{1}{2}\frac{\partial^2}{\partial x_j^2}g(x)+f(x)^{-1}\frac{\partial}{\partial x_j}g(x)\frac{\partial}{\partial x_j}f(x).$$

3.3 Mean Squared Error

The AMSE of the NW estimator $\hat{g}\left(x\right)$ is

$$AMSE\left(\hat{g}(x)\right)=\kappa_2^2\left(\sum_{j=1}^{q}h_j^2 B_j(x)\right)^2+\frac{R(k)^q\sigma^2(x)}{n\left|H\right|f(x)}$$

A weighted integrated MSE takes the form

$$\begin{aligned}
WIMSE &= \int AMSE\left(\hat{g}(x)\right)f(x)M(x)\left(dx\right) \\
&= \kappa_2^2\int\left(\sum_{j=1}^{q}h_j^2 B_j(x)\right)^2 f(x)M(x)\left(dx\right)+\frac{R(k)^q\int\sigma^2(x)M(x)dx}{nh_1h_2\cdots h_q}
\end{aligned}$$

where $M(x)$ is a weight function. Possible choices include $M(x)=f(x)$ and $M(x)=1\left(f(x)\geq\delta\right)$ for some $\delta>0$. The AMSE nees the weighting otherwise the integral will not exist.

3.4 Observations about the Asymptotic Distribution

In univariate regression, the optimal rate for the bandwidth is $h_0=Cn^{-1/5}$ with mean-squared convergence $O(n^{-2/5})$. In the multiple regressor case, the optimal bandwidths are $h_j=Cn^{-1/(q+4)}$ with convergence rate $O\left(n^{-2/(q+4)}\right)$. This is the same as for univariate and q-variate density estimation.

If higher-order kernels are used, the optimal bandwidth and convergence rates are again the same as for density estimation.

The asymptotic distribution depends on the kernel through $R(k)$ and κ_2. The optimal kernel minimizes $R(k)$, the same as for density estimation. Thus the Epanechnikov family is optimal for regression.

As the WIMSE depends on the first and second derivatives of the mean function $g(x)$, the optimal bandwidth will depend on these values. When the derivative functions $B_j(x)$ are larger, the optimal bandwidths are smaller, to capture the fluctuations in the function $g(x)$. When the derivatives are smaller, optimal bandwidths are larger, smoother more, and thus reducing the estimation variance.

For nonparametric regression, reference bandwidths are not natural. This is because there is no natural reference $g(x)$ which dictates the first and second derivative. Many authors use the rule-of-thumb bandwidth for density estimation (for the regressors X_i) but there is absolutely no justification for this choice. The theory shows that the optimal bandwidth depends on the curvature in the conditional mean $g(x)$, and this is independent of the marginal density $f(x)$ for which the rule-of-thumb is designed.

3.5 Limitations of the NW estimator

Suppose that $q = 1$ and the true conditional mean is linear $g(x) = \alpha + x\beta$. As this is a very simple situation, we might expect that a nonparametric estimator will work reasonably well. This is not necessarily the case with the NW estimator.

Take the absolutely simplest case that there is not regression error, i.e. $y_i = \alpha + X_i\beta$ identically. A simple scatter plot would reveal the deterministic relationship. How will NW perform?

The answer depends on the marginal distribution of the x_i. If they are not spaced at uniform distances, then $\hat{g}(x) \neq g(x)$. The NW estimator applied to purely linear data yields a nonlinear output!

One way to see the source of the problem is to consider the problem of nonparametrically estimating $E\left(X_i - x \mid X_i = x\right) = 0$. The numerator of the NW estimator of the expectation is

$$\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)(X_i - x)$$

but this is (generally) non-zero.

Can the problem by resolved by choice of bandwidth? Actually, it can make things worse. As the bandwidth increases (to increase smoothing) then $\hat{g}(x)$ collapses to a flat function. Recall that the NW estimator is also called the local constant estimator. It is approximating the regression function by a (local) constant. As smoothing increases, the estimator simplifies to a constant, not to a linear function.

Another limitation of the NW estimator occurs at the edges of the support. Again consider the case $q = 1$. For a value of $x \leq \min\left(X_i\right)$, then the NW estimator $\hat{g}(x)$ is an average only of y_i values for obsevations to the right of x. If $g(x)$ is positively sloped, the NW estimator will be upward biased. In fact, the estimator is inconsistent at the boundary. This effectively restricts application of the NW estimator to values of x in the interior of the support of the regressors, and this may too limiting.

3.6 Local Linear Estimator

We started this chapter by motivating the NW estimator at x by taking an average of the y_i values for observations such that X_i are in a neighborhood of x. This is a local constant approximation. Instead, we could fit a linear regression line through the observations in the same neighborhood. If we use a weighting function, this is called the local linear (LL) estimator, and it is quite popular in the recent nonparametric regression literature.

The idea is to fit the local model

$$y_i = \alpha + \beta' (X_i - x) + e_i$$

The reason for using the regressor $X_i - x$ rather than X_i is so that the intercept equals $g(x) = E(y_i \mid X_i = x)$. Once we get the estimates $\hat{\alpha}(x)$, $\hat{\beta}(x)$, we then set $\hat{g}(x) = \hat{\alpha}(x)$. Furthermore, we can use $\hat{\beta}(x)$ to estimate of $\frac{\partial}{\partial x} g(x)$.

If we simply fit a linear regression through observations such that $|X_i - x| \leq h$, this can be written as

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \left(y_i - \alpha - \beta' (X_i - x)\right)^2 1\left(|X_i - x| \leq h\right)$$

or setting

$$Z_i = \begin{pmatrix} 1 \\ X_i - x \end{pmatrix}$$

we have the explicit expression

$$\begin{aligned} \begin{pmatrix} \hat{\alpha}(x) \\ \hat{\beta}(x) \end{pmatrix} &= \left(\sum_{i=1}^{n} 1\left(|X_i - x| \leq h\right) Z_i Z_i' \right)^{-1} \left(\sum_{i=1}^{n} 1\left(|X_i - x| \leq h\right) Z_i y_i \right) \\ &= \left(\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) Z_i Z_i' \right)^{-1} \left(\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) Z_i y_i \right) \end{aligned}$$

where the second line is valid for any (multivariate) kernel funtion. This is a (locally) weighted regression of y_i on X_i. Algebraically, this equals a WLS estimator.

In contrast to the NW estimator, the LL estimator preserves linear data. That is, if the true data lie on a line $y_i = \alpha + X_i'\beta$, then for any sub-sample, a local linear regression fits exactly, so $\hat{g}(x) = g(x)$. In fact, we will see that the distribution of the LL estimator is invariant to the first derivative of g. It has zero bias when the true regression is linear.

As $h \to \infty$ (smoothing is increased), the LL estimator collapses to the OLS regression of y_i on X_i. In this sense LL is a natural nonparametric generalization of least-squares regression.

The LL estimator also has much better properties at the boundard than the NW estimator. Intuitively, even if x is at the boundard of the regression support, as the local linear estimator fits a (weighted) least-squares line through data near the boundary, if the true relationship is linear this estimator will be unbiased.

Deriving the asymptotic distribution of the LL estimator is similar to that of the NW estimator, but much more involved, so I will not present the argument here. It has the following asymptotic distribution. Let $\hat{g}(x) = \hat{\alpha}(x)$. Then

$$\sqrt{n\left|H\right|}\left(\hat{g}(x) - g(x) - \kappa_2 \sum_{j=1}^{q} h_j^2 \frac{1}{2}\frac{\partial^2}{\partial x_j^2} g(x)\right) \xrightarrow{d} N\left(0, \frac{R(k)^q \sigma^2(x)}{f(x)}\right)$$

This is quite similar to the distribution for the NW estimator, with one important difference that the bias term has been simplified. The term involving $f(x)^{-1}\frac{\partial}{\partial x_j}g(x)\frac{\partial}{\partial x_j}f(x)$ has been eliminated. The asymptotic variance is unchanged.

Strictly speaking, we cannot rank the AMSE of the NW versus the LL estimator. While a bias term has been eliminated, it is possible that the two terms have opposite signs and thereby cancel somewhat. However, the standard intuition is that a simplified bias term suggests reduced bias in practice. The AMSE of the LL estimator only depends on the second derivative of $g(x)$, while that of the NW estimator also depends on the first derivative. We expect this to translate into reduced bias.

Magically, this does not come as a cost in the asymptotic variance. These facts have led the statistics literature to focus on the LL estimator as the preferred approach.

While I agree with this general view, a side not of caution is warrented. Simple simulation experiments show that the LL estimator does not always beat the NW estimator. When the regression function $g(x)$ is quite flat, the NW estimator does better. When the regression function is steeper and curvier, the LL estimator tends to do better. The explanation is that while the two have identical asymptotic variance formulae, in finite samples the NW estimator tends to have a smaller variance. This gives it an advantage in contexts where estimation bias is low (such as when the regression function is flat). The reason why I mention this is that in many economic contexts, it is believed that the regression function may be quite flat with respect to many regressors. In this context it may be better to use NW rather than LL.

3.7 Local Polynomial Estimation

If LL improves on NW, why not local polynomial? The intuition is quite straightforward. Rather than fitting a local linear equation, we can fit a local quadratic, cubic, or polynomial of arbitrary order.

Let p denote the order of the local polynomial. Thus $p = 0$ is the NW estimator, $p = 1$ is the LL estimator, and $p = 2$ is a local quadratic.

Interestingly, the asymptotic behavior differs depending on whether p is even or odd.

When p is odd (e.g. LL), then the bias is of order $O(h^{p+1})$ and is proportional to $g^{(p+1)}(x)$

When p is even (e.g. NW or local quadratic), then the bias is of order $O(h^{p+2})$ but is proportional to $g^{(p+2)}(x)$ and $g^{(p+1)}(x)f^{(1)}(x)/f(x)$.

In either case, the variance is $O\left(\frac{1}{n\,|H|}\right)$

What happens is that by increasing the polynomial order from even to the next odd number, the order of the bias does change, but the bias simplifies.

By increasing the polynomial order from odd to the next even number, the bias order decreases. This effect is analogous to the bias reduction achieved by higher-order kernels.

While local linear estimation is gaining popularity in econometric practice, local polynomial methods are not typically used. I believe this is mostly because typical econometric applications have $q > 1$, and it is difficult to apply polynomial methods in this context.

3.8 Weighted Nadaraya-Watson Estimator

In the context of conditional distribution estimation, Hall et. al. (1999, JASA) and Cai (2002, ET) proposed a weighted NW estimator with the same asymptotic distribution as the LL estimator. This is discussed on pp. 187-188 of Li-Racine.

The estimator takes the form

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} p_i(x) K\left(H^{-1}(X_i - x)\right) y_i}{\sum_{i=1}^{n} p_i(x) K\left(H^{-1}(X_i - x)\right)}$$

where $p_i(x)$ are weights. The weights satisfy

$$\begin{aligned} p_i(x) &\geq 0 \\ \sum_{i=1}^{n} p_i(x) &= 1 \\ \sum_{i=1}^{n} p_i(x) K\left(H^{-1}(X_i - x)\right)(X_i - x) &= 0 \end{aligned}$$

The first two requirements set up the $p_i(x)$ as weights. The third equality requires the weights to force the kernel function to satisfy local linearity.

The weights are determined by empirical likelihood. Specifically, for each x, you maximize $\sum_{i=1}^{n} \ln p_i(x)$ subject to the above constraints. The solutions take the form

$$p_i(x) = \frac{1}{n\left(1 + \lambda'(X_i - x) K\left(H^{-1}(X_i - x)\right)\right)}$$

where λ is a Lagrange multiplier and is found by numerical optimization. For details about empirical likelihood, see my *Econometrics* lecture notes.

The above authors show that the estimator $\hat{g}(x)$ has the same asymptotic distribution as LL.

When the dependent variable is non-negative, $y_i \geq 0$, the standard and weighted NW estimators also satisfy $\hat{g}(x) \geq 0$. This is an advantage since it is obvious in this case that $g(x) \geq 0$. In contrast, the LL estimator is not necessarily non-negative.

An important disadvantage of the weighted NW estimator is that it is considerably more computationally cumbersome than the LL estimator. The EL weights must be found separately for each x at which $\hat{g}(x)$ is calculated.

3.9 Residual and Fit

Given any nonparametric estimator $\hat{g}(x)$ we can define the residual $\hat{e}_i = y_i - \hat{g}(X_i)$. Numerically, this requires computing the regression estimate at each observation. For example, in the case of NW estimation,

$$\hat{e}_i = y_i - \frac{\sum_{j=1}^n K\left(H^{-1}(X_j - X_i)\right) y_j}{\sum_{j=1}^n K\left(H^{-1}(X_j - X_i)\right)}$$

From $\hat{e}_i$ we can compute many conventional regression statistics. For example, the residual variance estimate is $n^{-1}\sum_{i=1}^n \hat{e}_i^2$, and R^2 has the standard formula.

One cautionary remark is that since the convergence rate for $\hat{g}$ is slower than $n^{-1/2}$, the same is true for many statistics computed from $\hat{e}_i$.

We can also compute the leave-one-out residuals

$$\begin{aligned} \hat{e}_{i,i-1} &= y_i - \hat{g}_{-i}(X_i) \\ &= y_i - \frac{\sum_{j \neq i} K\left(H^{-1}(X_j - X_i)\right) y_j}{\sum_{j \neq i} K\left(H^{-1}(X_j - X_i)\right)} \end{aligned}$$

3.10 Cross-Validation

For NW, LL and local polynomial regression, it is critical to have a reliable data-dependent rule for bandwidth selection. One popular and practical approach is cross-validation. The motivation starts by considering the sum-of-squared errors $\sum_{i=1}^n \hat{e}_i^2$. One could think about picking h to minimize this quantity. But this is analogous to picking the number of regressors in least-squares by minimizing the sum-of-squared errors. In that context the solution is to pick all possible regressors, as the sum-of-squared errors is monotonically decreasing in the number of regressors. The same is true in nonparametric regression. As the bandwidth h decreases, the in-sample "fit" of the model improves and $\sum_{i=1}^n \hat{e}_i^2$ decreases. As h shrinks to zero, $\hat{g}(X_i)$ collapses on y_i to obtain perfect fit, $\hat{e}_i$ shrinks to zero and so does $\sum_{i=1}^n \hat{e}_i^2$. It is clearly a poor choice to pick h based on this criterion.

Instead, we can consider the sum-of-squared leave-one-out residuals $\sum_{i=1}^n \hat{e}_{i,i-1}^2$. This is a reasonable criterion. Because the quality of $\hat{g}(X_i)$ can be quite poor for tail values of X_i, it may be more sensible to use a trimmed verion of the sum of squared residuals, and this is called the cross-validation criterion

$$CV(h) = \frac{1}{n}\sum_{i=1}^n \hat{e}_{i,i-1}^2 M(X_i)$$

(We have also divided by sample size for convenience.) The funtion $M(x)$ is a trimming function, the same as introduced in the definition of WIMSE earlier.

The cross-validation bandwidth h is that which minimizes $CV(h)$. As in the case of density estimation, this needs to be done numerically.

To see that the CV criterion is sensible, let us calculate its expectation. Since $y_i = g(X_i) + e_i$,

$$\begin{aligned} E(CV(h)) &= E\left((e_i + g(X_i) - \hat{g}_{-i}(X_i))^2 M(X_i)\right) \\ &= E\left((g(X_i) - \hat{g}_{-i}(X_i))^2 M(X_i)\right) - 2E(e_i(g(X_i) - \hat{g}_{-i}(X_i)) M(X_i)) + E\left(e_i^2 M(X_i)\right). \end{aligned}$$

The third term does not depend on the bandwidth so can be ignored. For the second term we use the law of iterated expectations, conditioning on X_i and I_{-i} (the sample excluding the i'th observation) to obtain

$$E(e_i(g(X_i) - \hat{g}_{-i}(X_i)) M(X_i) \mid I_{-i}, X_i) = E(e_i \mid X_i)(g(X_i) - \hat{g}_{-i}(X_i)) M(X_i) = 0$$

so the unconditional expecation is zero. For the first term we take expectations conditional on I_{-i} to obtain

$$E\left((g(X_i) - \hat{g}_{-i}(X_i))^2 M(X_i) \mid I_{-i}\right) = \int (g(x) - \hat{g}_{-i}(x))^2 M(x) f(x)(dx)$$

and thus the unconditional expectation is

$$\begin{aligned} E\left((g(X_i) - \hat{g}_{-i}(X_i))^2 M(X_i)\right) &= \int E(g(x) - \hat{g}_{-i}(x))^2 M(x) f(x)(dx) \\ &= \int E(g(x) - \hat{g}(x))^2 M(x) f(x)(dx) \\ &= \int MSE(\hat{g}(x)) M(x) f(x)(dx) \end{aligned}$$

which is $WIMSE(h)$.

We have shown that

$$E(CV(h)) = WIMSE(h) + E\left(e_i^2 M(X_i)\right)$$

Thus CV is an estimator of the weighted integrated squared error.

As in the case of density estimation, it can be shown that it is a good estimator of $WIMSE(h)$, in the sense that the minimizer of $CV(h)$ is consistent for the minimizer of $WIMSE(h)$. This holds true for NW, LL and other nonparametric methods. In this sense, cross-validation is a general, practical method for bandwidth selection.

3.11 Displaying Estimates and Pointwise Confidence Bands

When $q = 1$ it is simple to display $\hat{g}(x)$ as a function of x, by calculating the estimator on a grid of values.

When $q > 1$ it is less simple. Writing the estimator as $\hat{g}(x_1, x_2, ..., x_q)$, you can display it as a function of one variable, holding the others fixed. The variables held fixed can be set at their sample means, or varied across a few representative values.

When displaying an estimated regression function, it is good to include confidence bands. Typically this are pointwise confidence intervals, and can be computed using the $\hat{g}(x) \pm 2s(x)$ method, where $s(x)$ is a standard error. Recall that the asymptotic distribution of the NW and LL estimators take the form

$$\sqrt{nh_1 \cdots h_q}\left(\hat{g}(x) - g(x) - Bias(x)\right) \xrightarrow{d} N\left(0, \frac{R(k)^q \sigma^2(x)}{f(x)}\right).$$

Ignoring the bias (as it cannot be estimated well), this suggests the standard error formula

$$s(x) = \sqrt{\frac{R(k)^q \hat{\sigma}^2(x)}{nh_1 \cdots h_q \hat{f}(x)}}$$

where $\hat{f}(x)$ is an estimate of $f(x)$ and $\hat{\sigma}^2(x)$ is an estimate of $\sigma^2(x) = E\left(e_i^2 \mid X_i = x\right)$.

A simple choice for $\hat{\sigma}^2(x)$ is the sample mean of the residuals $\hat{\sigma}^2$. But this is valid only under conditional homoskedasticity. We discuss nonparametric estimates for $\sigma^2(x)$ shortly.

3.12 Uniform Convergence

For theoretical purposes we often need nonparametric estimators such as $\hat{f}(x)$ or $\hat{g}(x)$ to converge uniformly. The primary applications are two-step and semiparametric estimators which depend on the first step nonparametric estimator. For example, if a two-step estimator depends on the residual $\hat{e}_i$, we note that

$$\hat{e}_i - e_i = g\left(X_i\right) - \hat{g}\left(X_i\right)$$

is hard to handle (in terms of stochastically bounding), as it is an estimated function evaluated at a random variable. If $\hat{g}(x)$ converges to $g(x)$ pointwise in x, but not uniformly in x, then we don't know if the difference $g\left(X_i\right) - \hat{g}\left(X_i\right)$ is converging to zero or not. One solution is to apply a uniform convergence result. That is, the above expression is bounded in absolute value, for $|X_i| \le C$ for some $C < \infty$ by

$$\sup_{|x| \le C} \left|g\left(x\right) - \hat{g}\left(x\right)\right|$$

and this is the object of study for uniform convergence.

It turns out that there is some cost to obtain uniformity. While the NW and LL estimators pointwise converge at the rates $n^{-2/(q+4)}$ (the square root of the MSE convergence rate), the uniform convergence rate is

$$\sup_{|x|\leq C}|g\left(x\right)-\hat{g}\left(x\right)|=O_{p}\left(\left(\frac{\ln n}{n}\right)^{2/(q+4)}\right)$$

The $O_p\left(\cdot\right)$ symbol means "bounded in probability", meaning that the LHS is bounded beneath a constant this the rate, with probability arbitrarily close to one. Alternatively, the same rate holds almost surely. The difference with the pointwise case is the addition of the extra $\ln n$ term. This is a very slow penalty, but it is a penalty none-the-less.

This rate was shown by Stein to be the best possible rate, so the penalty is not an artifact of the proof technique.

A recent paper of mine provides some generalizations of this result, allowing for dependent data (time series). B. Hansen, Econometric Theory, 2008.

One important feature of this type of bound is the restriction of x to the compact set $|x|\leq C$. This is a bit unfortunate as in applications we often want to apply uniform convergence over the entire support of the regressors, and the latter can be unbounded. One solution is to ignore this technicality, and just "assume" that the regressors are bounded. Another solution is to apply the result using "trimming", a technique which we will probably discuss later, when we do semiparametrics. Finally, as shown in my 2008 paper, it is also possible to allow the constant $C=C_n$ to diverge with n, but at the cost of slowing down the rate of convergence on the RHS.

3.13 NonParametric Variance Estimation

Let $\sigma^2(x)=var\left(y_i\mid X_i=x\right)$. It is sometimes of direct economic interest to estimate $\sigma^2(x)$. In other cases we just want to estimate it to get a confidence interval for $g(x)$.

The following method is recommended. Write the model as

$$\begin{aligned}
y_i &= g\left(X_i\right)+e_i\\
E\left(e_i\mid X_i\right) &= 0\\
e_i^2 &= \sigma^2\left(X_i\right)+\eta_i\\
E\left(\eta_i\mid X_i\right) &= 0
\end{aligned}$$

Then $\sigma^2(x)$ is the regression function of e_i^2 on X_i.

If e_i^2 were observed, this could be done using NW, weighted NW, or LL regression. While e_i^2 is not observed, it can be replaced by $\hat{e}_i^2$ where $\hat{e}_i=y_i-\hat{g}\left(X\right)$ are the nonparametric regression residuals. Using a NW estimator

$$\hat{\sigma}^2(x)=\frac{\sum_{i=1}^{n}K\left(H^{-1}\left(X_i-x\right)\right)\hat{e}_i^2}{\sum_{i=1}^{n}K\left(H^{-1}\left(X_i-x\right)\right)}$$

and similarly using weighted NW or LL. The bandwidths H are not the same as for estimation of $\hat{g}(x)$, although we use the same notation.

As discussed earlier, the LL estimator $\hat{\sigma}^2(x)$ is not guarenteed to be non-negative, while the NW and weighted NW estimators are always non-negative (if non-negative kernels are used).

Fan and Yao (1998, Biometrika) analyze the asymptotic distribution of this estimator. They obtain the surprising result that the asymptotic distribution of this two-step estimator is identical to that of the one-step idealized estimator

$$\tilde{\sigma}^2(x) = \frac{\sum_{i=1}^n K\left(H^{-1}(X_i - x)\right) e_i^2}{\sum_{i=1}^n K\left(H^{-1}(X_i - x)\right)}.$$

That is, the nonparametric regression of $\hat{e}_i^2$ on x_i is asymptotically equivalent to the nonparametric regression of e_i^2 on x_i.

Technically, they demonstrated this result when $\hat{g}$ and $\hat{\sigma}^2$ are computed using LL, but from the nature of the argument it appears that the same holds for the NW estimator. They also only demonstrated the result for $q = 1$, but the extends to the $q > 1$ case.

This is a neat result, and is not typical in two-step estimation. One convenient implication is that we can pick bandwidths in each step based on conventional one-step regression methods, ignoring the two-step nature of the problem. Additionally, we do not have to worry about the first-step estimation of $g(x)$ when computing confidence intervals for $\sigma^2(x)$.

4 Conditional Distribution Estimation

4.1 Estimators

The conditional distribution (CDF) of y_i given $X_i = x$ is

$$\begin{aligned} F(y \mid x) &= P(y_i \leq y \mid X_i = x) \\ &= E(1(y_i \leq y) \mid X_i = x). \end{aligned}$$

This is the conditional mean of the random variable $1(y_i \leq y)$. Thus the CDF is a regression, and can be estimated using regression methods.

One difference is that $1(y_i \leq y)$ is a function of the argument y, so CDF estimation is a set of regressions, one for each value of y.

Standard CDF estimators include the NW, LL, and WNW. The NW can be written as

$$\hat{F}(y \mid x) = \frac{\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) 1(y_i \leq y)}{\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right)}$$

The NW and WNW estimators have the advantages that they are non-negative and non-decreasing in y, and are thus valid CDFs.

The LL estimator does not necessarily satisfy these properties. It can be negative, and need not be monotonic in y.

As we learned for regression estimation, the LL and WMW estimators both have "better" bias and boundary properties. Putting these two observations together, it seems reasonable to consider using the WNW estimator.

The estimator $\hat{F}(y \mid x)$ is smooth in x, but a step function in y. We discuss later estimators which are smooth in y.

4.2 Asymptotic Distribution

Recall that in the case of kernel regression, we had

$$\sqrt{n\,|H|}\left(\hat{g}(x) - g(x) - \kappa_2 \sum_{j=1}^{q} h_j^2 B_j(x)\right) \xrightarrow{d} N\left(0, \frac{R(k)^q \sigma^2(x)}{f(x)}\right)$$

where $\sigma^2(x)$ was the conditional variance of the regression, and the $B_j(x)$ equals (for NW)

$$B_j(x) = \frac{1}{2}\frac{\partial^2}{\partial x_j^2} g(x) + f(x)^{-1} \frac{\partial}{\partial x_j} g(x) \frac{\partial}{\partial x_j} f(x)$$

while for LL and WNW the bias term is just the first part.

Clearly, for any fixed y, the same theory applies. In the case of CDF estimation, the regression equation is

$$1\left(y_i \leq y\right) = F\left(y \mid X_i\right) + e_i\left(y\right)$$

where $e_i(y)$ is conditionally mean zero and has conditional variance function

$$\sigma^2(x) = F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right).$$

(We know the conditional variance takes this form because the dependent variable is binary.) I write the error as a function of y to emphasize that it is different for each y.

In the case of LL or NWW, the bias terms are

$$B_j\left(y \mid x\right) = \frac{1}{2}\frac{\partial^2}{\partial x_j^2}F\left(y \mid x\right)$$

the curvature in the CDF with respect to the conditioning variables.

We thus find for all (y, x)

$$\sqrt{n\left|H\right|}\left(\hat{F}\left(y \mid x\right) - F\left(y \mid x\right) - \kappa_2\sum_{j=1}^{q}h_j^2 B_j\left(y \mid x\right)\right) \xrightarrow{d} N\left(0, \frac{R(k)^q F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right)}{f(x)}\right)$$

and

$$AMSE\left(\hat{F}\left(y \mid x\right)\right) = \kappa_2^2\left(\sum_{j=1}^{q}h_j^2 B_j\left(y \mid x\right)\right)^2 + \frac{R(k)^q F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right)}{n\left|H\right| f(x)}$$

In the $q = 1$ case

$$AMSE\left(\hat{F}\left(y \mid x\right)\right) = h^4\kappa_2^2 B\left(y \mid x\right)^2 + \frac{R(k)F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right)}{nhf(x)}.$$

In the regression case we defined the WIMSE as the integral of the AMSE, weighting by $f(x)M(x)$. Here we also integrate over y. For $q = 1$

$$\begin{aligned} WIMSE &= \int\int AMSE\left(\hat{F}\left(y \mid x\right)\right) f(x)M(x)\left(dx\right)dy \\ &= h^4\kappa_2^2\int\int B\left(y \mid x\right)^2 dy f(x)M(x)\left(dx\right) + \frac{R(k)\int\int F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right)dyM(x)dx}{nh} \end{aligned}$$

The integral over y does not need weighting since $F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right)$ declines to zero as y tends to either limit.

Observe that the converge rate is the same as in regression. The optimal bandwidths are the same rates as in regression.

4.3 Bandwidth Selection

I do not believe that bandwidth choice for nonparametric CDF estimation is widely studied. Li-Racine suggest using a CV method based on conditional density estimation.

It should also be possible to directly apply CV methods to CDF estimation.

The leave-one-out residuals are

$$\hat{e}_{i,i-1}(y) = 1(y_i \leq y) - \hat{F}_{-i}(y \mid X_i)$$

So the CV criterion for any fixed y is

$$\begin{aligned} CV(y,h) &= \frac{1}{n}\sum_{i=1}^{n} \hat{e}_{i,i-1}(y)^2 M(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \Big(1(y_i \leq y) - \hat{F}_{-i}(y \mid X_i)\Big)^2 M(X_i) \end{aligned}$$

If you wanted to estimate the CDF at a single value of y you could pick h to minimize this criterion.

For estimation of the entire function, we want to integrate over the values of y. One method is

$$\begin{aligned} CV(h) &= \int CV(y,h)dy \\ &\simeq \delta\sum_{j=1}^{N} CV(y_j^*,h) \end{aligned}$$

where y_j^* is a grid of values over the support of y_i such that $y_j - y_{j=1} = \delta$. To calculate this quantity, it involves N times the number of calculations as for regression, as the leave-one-out computations are done for each y_j^* on the grid. My guess is that the grid over the y values could be coarse, e.g. one could set $N = 10$.

4.4 Smoothed Distribution Estimators - Unconditional Case

The CDF estimators introduced above are not smooth, but are discontinuous step functions. For some applications this may be inconvenient. It may be desireable to have a smooth CDF estimate as an input for a semiparametric estimator. It is also the case that smoothing will improve high-order estimation efficiency. To see this, we need to return to the case of univariate data.

Recall that the univariate DF estimator for iid data y_i is

$$\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} 1(y_i \leq y)$$

It is easy to see that this estimator is unbiased and has variance $F(y)(1-F(y))/n$.

Now consider a smoothed estimator

$$\tilde{F}(y) = \frac{1}{n}\sum_{i=1}^{n} G\left(\frac{y-y_i}{h}\right)$$

where $G(x) = \int_{-\infty}^{x} k(u)du$ is a kernel distribution function (the integral of a univariate kernel function). Thus $\tilde{F}(y) = \int_{-\infty}^{y} \hat{f}(x)dx$ where $\hat{f}(x)$ is the kernel density estimate.

To calculate its expectation

$$\begin{aligned} \mathrm{E}\tilde{F}(y) &= \mathrm{E}G\left(\frac{y-y_i}{h}\right) \\ &= \int G\left(\frac{y-x}{h}\right) f(x)dx \\ &= h\int G\left(u\right) f\left(y-hu\right) du \end{aligned}$$

the last using the change of variables $u = (y-x)/h$ or $x = y - hu$ with Jacobian h.

Next, do not expand $f\left(y-hu\right)$ in a Taylor expansion, because the moments of G do not exist. Instead, first use integration by parts. The integral of f is F and that of $hf\left(y-hu\right)$ is $-F\left(y-hu\right)$, and the derivative of $G(u)$ is $k(u)$. Thus the above equals

$$\int k\left(u\right) F\left(y-hu\right) du$$

which can now be expanded using Taylor's expansion, yielding

$$\mathrm{E}\tilde{F}(y) = F(y) + \frac{1}{2}\kappa_2 h^2 f^{(1)}(y) + o\left(h^2\right)$$

Just as in other estimation contexts, we see that the bias of $\tilde{F}(y)$ is of order h^2, and is proportional to the second derivative of what we are estimating, as $F^{(2)}(y) = f^{(1)}(y)$

Thus smoothing introduces estimation bias.

The interesting part comes in the analysis of variance.

$$\begin{aligned} var\left(\tilde{F}(y)\right) &= \frac{1}{n} var\left(G\left(\frac{y-y_i}{h}\right)\right) \\ &= \frac{1}{n}\left(\mathrm{E}G\left(\frac{y-y_i}{h}\right)^2 - \left(\mathrm{E}G\left(\frac{y-y_i}{h}\right)\right)^2\right) \\ &\simeq \frac{1}{n}\left(\int G\left(\frac{y-x}{h}\right)^2 f(x)dx - F(y)^2\right) \end{aligned}$$

Let's calculate this integral. By a change of variables

$$\int G\left(\frac{y-x}{h}\right)^2 f(x)dx = h\int G\left(u\right)^2 f(y-hu)du$$

Once again we cannot direct apply a Taylor expansion, but need to first do integration-by-parts. Again the integral of $hf(y-hu)$ is $-F(y-hu)$. The derivative of $G(u)^2$ is $2G(u)k(u)$. So the above is

$$2\int G(u)\,k(u)F(y-hu)\,du$$

and then applying a Taylor expansion, we obtain

$$F(y)\left(2\int G(u)\,k(u)du\right)-f(y)\,h\left(2\int G(u)\,k(u)udu\right)+o(h)$$

since $F^{(1)}(y)=f(y)$.

Now since the derivative of $G(u)^2$ is $2G(u)k(u)$, it follows that the integral of $2G(u)k(u)$ is $G(u)^2$, and thus the first integral over $(-\infty,\infty)$ is $G(\infty)^2-G(-\infty)^2=1-0=1$ since $G(u)$ is a distribution function. Thus the first part is simply $F(y)$. Define

$$\alpha(k)=2\int G(u)\,k(u)udu>0$$

For any symmetric kernel k, $\alpha(k)>0$. This is because for $u>0$, $G(u)>G(-u)$, thus

$$\int_0^\infty G(u)\,k(u)udu>\int_0^\infty G(-u)\,k(u)udu=-\int_{-\infty}^0 G(u)\,k(u)udu$$

and so the integral over $(-\infty,\infty)$ is positive. Integrated kernels and the value $\alpha(k)$ are given in the following table.

Kernel	**Integrated Kernel**	$\alpha(k)$
Epanechnikov	$G_1(u)=\frac{1}{4}\left(2+3u-u^3\right)1\left(\lvert u\rvert\le 1\right)$	$9/35$
Biweight	$G_2(u)=16\left(8+15u-10u^3+3u^5\right)1\left(\lvert u\rvert\le 1\right)$	$50/231$
Triweight	$G_3(u)=32\left(16+35u-35u^2+21u^5-5u^7\right)1\left(\lvert u\rvert\le 1\right)$	$245/1287$
Gaussian	$G_\phi(u)=\Phi(u)$	$1/\sqrt{\pi}$

Together, we have

$$\begin{aligned}
var\left(\tilde{F}(y)\right) &\simeq \frac{1}{n}\left(\int G\left(\frac{y-x}{h}\right)^2 f(x)dx-F(y)^2\right)\\
&= \frac{1}{n}\left(F(y)-F(y)^2-\alpha(k)\,f(y)\,h+o(h)\right)\\
&= \frac{F(y)\left(1-F(y)\right)}{n}-\alpha(k)\,f(y)\,\frac{h}{n}+o\left(\frac{h}{n}\right)
\end{aligned}$$

The first part is the variance of $\hat{F}(x)$, the unsmoothed estimator. Smoothing reduces the variance by $\alpha_0 f(y)\,\frac{h}{n}$.

Its MSE is

$$MSE\left(\tilde{F}(y)\right) = \frac{F(y)\,(1-F(y))}{n} - \alpha\,(k)\,f\,(y)\,\frac{h}{n} + \frac{\kappa_2^2 h^4}{4} f^{(1)}(y)^2$$

The integrated MSE is

$$\begin{aligned} MISE\left(\tilde{F}(y)\right) &= \int MSE\left(\tilde{F}(y)\right) dy \\ &= \frac{\int F(y)\,(1-F(y))\,dy}{n} - \alpha\,(k)\,\frac{h}{n} + \frac{\kappa_2^2 h^4 R\left(f^{(1)}\right)}{4} \end{aligned}$$

where

$$R\left(f^{(1)}\right) = \int f^{(1)}(y)^2 dy$$

The first term is independent of the smoothing parameter h (and corresponds to the integrated variance of the unsmoothed EDF estimator). To find the optimal bandwidth, take the FOC:

$$\frac{d}{dh} MISE\left(\hat{F}(y)\right) = -\frac{\alpha\,(k)}{n} + \kappa_2^2 h^3 R\left(f^{(1)}\right) = 0$$

and solve to find

$$h_0 = \left(\frac{\alpha\,(k)}{\kappa_2^2 R\left(f^{(1)}\right)}\right)^{1/3} n^{-1/3}$$

The optimal bandwidth converges to zero at the fast $n^{-1/3}$ rate.

Does smoothing help? The unsmoothed estimator has MISE of order n^{-1}, and the smoothed estimator (with optimal bandwidth) is of order $n^{-1} - n^{-4/3}$. We can thus think of the gain in the scaled MISE as being of order $n^{-4/3}$, which is of smaller order than the original n^{-1} rate.

It is important that the bandwidth not be too large. Suppose you set $h \propto n^{-1/5}$ as for density estimation. Then the square bias term is of order $h^4 \propto n^{-4/5}$ which is larger than the leading term. In this case the smoothed estimator has larger MSE than the usual estimator! Indeed, you need h to be of smaller order than $n^{-1/4}$ for the MSE to be no worse than the unusual case.

For practical bandwidth selection, Li-Racine and Bowman et. al. (1998) recommend a CV method. For fixed y the criterion is

$$CV(h,y) = \frac{1}{n}\sum_{i=1}^{n}\left(1\,(y_i \leq y) - \tilde{F}_{-i}\,(y)\right)^2$$

which is the sum of squared leave-one-out residuals. For a global estimate the criterion is

$$CV(h) = \int CV(h,y)dy$$

and this can be approximated by a summation over a grid of values for y.

This is essentially the same as the CV criterion we introduced above in the conditional case.

4.5 Smoothed Conditional Distribution Estimators

The smoothed versions of the CDF estimators replace the indicator functions $1\left(y_i \leq y\right)$ with the integrated kernel $G\left(\frac{y-y_i}{h_0}\right)$ where we will use h_0 to denote the bandwidth smoothing in the y direction.

The NW version is

$$\tilde{F}\left(y \mid x\right)=\frac{\sum_{i=1}^{n} K\left(H^{-1}\left(X_i-x\right)\right) G\left(\frac{y-y_i}{h_0}\right)}{\sum_{i=1}^{n} K\left(H^{-1}\left(X_i-x\right)\right)}$$

with $H=\{h_1, ... h_q\}$. The LL is obtained by a local linear regression of $G\left(\frac{y-y_i}{h_0}\right)$ on X_i-x with bandwidths H. And similarly the WNW.

What is its distribution? It is essentially that of $\hat{F}\left(y \mid x\right)$, plus an additional bias term, minus a variance term.

First take bias. Recall

$$Bias\left(\hat{F}\left(y \mid x\right)\right) \simeq \kappa_2 \sum_{j=1}^{q} h_j^2 B_j\left(y \mid x\right)$$

where for LL and WNW

$$B_j\left(y \mid x\right)=\frac{1}{2} \frac{\partial^2}{\partial x_j^2} F\left(y \mid x\right).$$

And for smoothed DF estimation, the bias term is

$$\kappa_2 h^2 \frac{1}{2} \frac{\partial^2}{\partial y^2} F(y)$$

If you work out the bias of the smoothed CDF, you find it is the sum of these two, that is $\tilde{F}\left(y \mid x\right)$

$$Bias\left(\tilde{F}\left(y \mid x\right)\right) \simeq \kappa_2 \sum_{j=0}^{q} h_j^2 B_j\left(y \mid x\right)$$

where for $j \geq 1$ the $B_j\left(y \mid x\right)$ are the same as before, and for $j=0$,

$$B_0\left(y \mid x\right)=\frac{1}{2} \frac{\partial^2}{\partial y^2} F\left(y \mid x\right).$$

For variance, recall

$$var\left(\hat{F}\left(y \mid x\right)\right)=\frac{R(k)^q F\left(y \mid x\right)\left(1-F\left(y \mid x\right)\right)}{f(x) n\left|H\right|}$$

and for smoothed DF estimation, the variance was reduced by the term $-\alpha_0 f\left(y\right) \frac{h}{n}$. In the CDF case it turns out to be similarly adjusted:

$$var\left(\tilde{F}\left(y \mid x\right)\right) = \frac{R(k)^q\left[F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right) - h_0\alpha\left(k\right)f\left(y \mid x\right)\right]}{f(x)n\left|H\right|}$$

In sum, the MSE is

$$MSE\left(\tilde{F}\left(y \mid x\right)\right) = \kappa_2^2\left(\sum_{j=0}^{q} h_j^2 B_j\left(y \mid x\right)\right)^2 + \frac{R(k)^q\left[F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right) - h_0\alpha\left(k\right)f\left(y \mid x\right)\right]}{f(x)n\left|H\right|}$$

The WIMSE, $q = 1$ case, is

$$\begin{aligned} WIMSE &= \int\int AMSE\left(\tilde{F}\left(y \mid x\right)\right)f(x)M(x)\left(dx\right)dy \\ &= \kappa_2^2\int\int\left(h_0^2B_0\left(y \mid x\right) + h_1^2B_1\left(y \mid x\right)\right)^2 dyf(x)M(x)\left(dx\right) \\ &\quad + \frac{R(k)\left[\int\int F\left(y \mid x\right)\left(1 - F\left(y \mid x\right)\right)dyM(x)dx - h_0\alpha(k)\int M(x)dx\right]}{nh_1} \end{aligned}$$

4.6 Bandwidth Choice

First, consider the optimal bandwidth rates.

As smoothing in the y direction only affects the higher-order asymptotic distribution, it should be clear that the optimal rates for $h_1, ..., h_q$ is unchanged from the unsmoothed case, and is therefore equal to the regression setting. Thus the optimal bandwidth rates are $h_j \sim n^{-1/(4+q)}$ for $j \geq 1$.

Substituting these rates into the MSE equation, and ignoring constants, we have

$$MSE\left(\tilde{F}\left(y \mid x\right)\right) \sim \left(h_0^2 + n^{-2/(4+q)}\right)^2 + \frac{1}{n^{4/(4+q)}} - \frac{h_0}{n^{4/(4+q)}}$$

Differentiating with respect to h_0

$$0 = 4\left(h_0^2 + n^{-2/(4+q)}\right)h_0 - \frac{1}{n^{4/(4+q)}}$$

and since h_0 will be of smaller order than $n^{-1/(4+q)}$, we can ignore the h_0^3 term, and then solving the remainder we obtain $h_0 \sim n^{-2/(4+q)}$. E.g. for $q = 1$ then the optimal rate is $h_0 \sim n^{-2/5}$.

What is the gain from smoothing? With optimal bandwidth, the MISE is reduced by a term of order $n^{-6/(4+q)}$. This is $n^{-6/5}$ for $q = 1$ and n^{-1} for $q = 2$. This gain increases as q increases. Thus the gain in efficiency (from smoothing) is increased when X is of higher dimension. Intuitively, increasing X is equivalent to reducing the effective sample size, increasing the gain from smoothing.

How should the bandwidth be selected?

Li-Racine recommend picking the bandwidths by using a CV method for conditional density estimation, and then rescaling.

As an alternative, we can use CV directly for the CDF estimate. That is, define the CV criterion

$$\begin{aligned} CV(y,h) &= \frac{1}{n}\sum_{i=1}^{n}\Big(1\left(y_i \le y\right) - \tilde{F}_{-i}\left(y \mid X_i\right)\Big)^2 M\left(X_i\right) \\ CV(h) &= \int CV(h,y)dy \end{aligned}$$

where $h = (h_0, h_1, ..., h_q)$ includes smoothing in both the y and x directions. The estimator $\tilde{F}_{-i}$ is the smooth leave-one-out estimator of F. This formulae allows includes NW, LL and WNW estimation.

The second integral can be approximated using a grid.

To my knowledge, this procedure has not been formally investigated.

5 Conditional Density Estimation

5.1 Estimators

The conditional density of y_i given $X_i = x$ is $f(y \mid x) = f(y,x)/f(x)$. An natural estimator is

$$\begin{aligned}
\hat{f}(y \mid x) &= \frac{\hat{f}(y,x)}{\hat{f}(x)} \\
&= \frac{\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) k_{h_0}(y_i - y)}{\sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right)}
\end{aligned}$$

where $H = diag\{h_1, ..., h_q\}$ and $k_h(u) = h^{-1}k(u/h)$. This is the derivative of the smooth NW-type estimator $\tilde{F}(y \mid x)$. The bandwidth h_0 smooths in the y direction and the bandwidths $h_1, ..., h_q$ smooth in the X directions.

This is the NW estimator of the conditional mean of $Z_i = k_{h_0}(y - y_i)$ given $X_i = x$.

Notice that

$$\begin{aligned}
\mathrm{E}(Z_i \mid X_i = x) &= \int \frac{1}{h_0} k\left(\frac{v-y}{h_0}\right) f(v \mid x)\, dv \\
&= \int k(u) f(y - uh_0 \mid x)\, du \\
&\simeq f(y \mid x) + \frac{h_0^2 \kappa_2}{2} \frac{\partial^2}{\partial y^2} f(y \mid x).
\end{aligned}$$

We can view conditional density estimation as a regression problem. In addition to NW, we can use LL and WNW estimation. The local polynomial method was proposed in a paper by Fan, Yao and Tong (Biometrika, 1996) and has been called the "double kernel" method.

5.2 Bias

By the formula for NW regression of Z_i on $X_i = x$,

$$\begin{aligned}
\mathrm{E}\hat{f}(y \mid x) &= \mathrm{E}(Z_i \mid X_i = x) + \kappa_2 \sum_{j=1}^{q} h_j^2 B_j(y \mid x) \\
&= f(y \mid x) + \kappa_2 \sum_{j=0}^{q} h_j^2 B_j(y \mid x)
\end{aligned}$$

where

$$\begin{aligned}
B_0(y \quad | \quad x) &= \frac{1}{2} \frac{\partial^2}{\partial^2 y} f(y \mid x) \\
B_j(y \quad | \quad x) &= \frac{1}{2} \frac{\partial^2}{\partial x_j^2} f(y \mid x) + f(x)^{-1} \frac{\partial}{\partial x_j} f(y \mid x) \frac{\partial}{\partial x_j} f(x), \qquad j > 0
\end{aligned}$$

as B_j are the curvature of $\mathrm{E}\left(Z_i \mid X_i = x\right) \simeq f\left(y \mid x\right)$ with respect to x_j. For LL or WNW

$$B_j(y \mid x) = \frac{1}{2}\frac{\partial^2}{\partial x_j^2} f\left(y \mid x\right), \qquad j > 0$$

The bias of $\hat{f}\left(y \mid x\right)$ for $f\left(y \mid x\right)$ is $\kappa_2 \sum_{j=0}^{q} h_j^2 B_j(y \mid x)$.

For the bias to converge to zero with n, all bandwidths must decline to zero.

5.3 Variance

By the formula for NW regression of Z_i on $X_i = x$,

$$var\left(\hat{f}\left(y \mid x\right)\right) \simeq \frac{R(k)^q}{nh_1 \cdots h_q f(x)} var\left(Z_i \mid X_i = x\right)$$

We calculate that

$$\begin{aligned} var\left(Z_i \mid X_i = x\right) &= \mathrm{E}\left(Z_i^2 \mid X_i = x\right) - \left(\mathrm{E}\left(Z_i \mid X_i = x\right)\right)^2 \\ &\simeq \frac{1}{h_0^2}\int k\left(\frac{v-y}{h_0}\right)^2 f\left(v \mid x\right) dv \\ &= \frac{1}{h_0}\int k\left(u\right)^2 f\left(y - uh \mid x\right) du \\ &\simeq \frac{R(k) f\left(y \mid x\right)}{h_0}. \end{aligned}$$

Substituting this into the expression for the estimation variance,

$$\begin{aligned} var\left(\hat{f}\left(y \mid x\right)\right) &\simeq \frac{R(k)^q}{nh_1 \cdots h_q f(x)} var\left(Z_i \mid X_i = x\right) \\ &= \frac{R(k)^{q+1} f\left(y \mid x\right)}{nh_0 h_1 \cdots h_q f(x)} \end{aligned}$$

What is key is that the variance of the conditional density depends inversely upon all bandwidths.

For the variance to tend to zero, we thus need $nh_0 h_1 \cdots h_q \to \infty$.

5.4 MSE

$$AMSE\left(\hat{f}\left(y \mid x\right)\right) = \kappa_2^2 \left(\sum_{j=0}^{q} h_j^2 B_j(y \mid x)\right)^2 + \frac{R(k)^{q+1} f\left(y \mid x\right)}{nh_0 h_1 \cdots h_q f(x)}$$

In this problem, the bandwidths enter symmetrically. Thus the optimal rates for h_0 and the other bandwidths will be equal. To see this, let h be a common bandwidth and ignoring constants, then

$$AMSE\left(\hat{f}\left(y \mid x\right)\right) \sim h^4 + \frac{1}{nh^{1+q}}$$

with optimal solution

$$h \sim n^{-1/(5+q)}.$$

Thus if $q = 1$, $h \sim n^{-1/6}$ or $q = 2$, $h \sim n^{-1/7}$. This is the same rate as for multivariate density estimation (estimation of the joint density $f(y, x)$). The resulting convergence rate for the estimator is the same as multivariate density estimation.

5.5 Cross-validation

Fan and Yim (2004, Biometrika) and Hall, Racine and Li (2004) have proposed a cross-validation method appropriate for nonparametric conditional density estimators. In this section we describe this method and its application to our estimators. For an estimator $\hat{f}(y \mid x)$ of $f(y \mid x)$ define the integrated squared error

$$\begin{aligned}
I(h) &= \int\int \left(\tilde{f}(y \mid x) - f(y \mid x)\right)^2 M(x)f(x)dydx \\
&= \int\int \tilde{f}(y \mid x)^2 M(x)f(x)dydx - 2\int\int \tilde{f}(y \mid x) M(x) f(y \mid x) f(x)dydx + \int\int f(y \mid x)^2 M(x)f(x)dydx \\
&= \mathrm{E}\left(\int \tilde{f}(y \mid X_i)^2 M(X_i)\right) - 2\mathrm{E}\left(\tilde{f}(y_i \mid x_i) M(X_i)\right) + \mathrm{E}\left(\int f(y \mid X_i)^2 M(X_i)dy\right) \\
&= I_1(h) - 2I_2(h) + I_3.
\end{aligned}$$

Note that I_3 does not depend on the bandwidths and is thus irrelvant.

Let $\hat{f}_{-i}(y \mid X_i)$ denote the estimator $\hat{f}(y \mid x)$ at $x = X_i$ with observation i omitted. For the NW estimator this equals

$$\hat{f}_{-i}(y \mid X_i) = \frac{\sum_{j \neq i} K\left(H^{-1}(X_i - X_j)\right) k_{h_0}(y_i - y)}{\sum_{j \neq i} K\left(H^{-1}(X_i - X_j)\right)}$$

The cross-validation estimators of I_1 and I_2 are

$$\begin{aligned}
\hat{I}_1(h) &= \frac{1}{n}\sum_{i=1}^{n} M(X_i) \int \tilde{f}_{-i}(y \mid X_i)^2 \, dy \\
\hat{I}_2(h) &= \frac{1}{n}\sum_{i=1}^{n} M(X_i)\tilde{f}_{-i}(Y_i \mid X_i).
\end{aligned}$$

Rhe cross-validation criterion is

$$CV(h) = \hat{I}_1(h) - 2\hat{I}_2(h).$$

The cross-validated bandwidths $h_0, h_1, ..., h_q$ are those which jointly minimize $CV(h)$

For the case of NW estimation

$$\begin{aligned}\hat{I}_1 &= \frac{1}{n}\sum_{i=1}^{n} M(X_i)\frac{\sum_{j\neq i}\sum_{k\neq i} K\left(H^{-1}\left(X_i - X_j\right)\right) K\left(H^{-1}\left(X_i - X_k\right)\right) \int k_{h_0}\left(y_j - y\right) k_{h_0}\left(y_k - y\right) dy}{\left(\sum_{j\neq i} K\left(H^{-1}\left(X_i - X_j\right)\right)\right)^2} \\ &= \frac{1}{n}\sum_{i=1}^{n} M(X_i)\frac{\sum_{j\neq i}\sum_{k\neq i} K\left(H^{-1}\left(X_i - X_j\right)\right) K\left(H^{-1}\left(X_i - X_k\right)\right) \bar{k}_{h_0}\left(y_i - y_j\right)}{\left(\sum_{j\neq i} K\left(H^{-1}\left(X_i - X_j\right)\right)\right)^2},\end{aligned}$$

where $\bar{k}$ is the convolution of k with itself, and

$$\hat{I}_2(h) = \frac{1}{n}\sum_{i=1}^{n} M(X_i)\frac{\sum\limits_{j\neq i} K\left(H^{-1}\left(X_i - X_j\right)\right) k_{h_0}\left(y_i - y_j\right)}{\sum\limits_{j\neq i} K\left(H^{-1}\left(X_i - X_j\right)\right)}.$$

5.1 Two-Step Conditional Density Estimator

We can write

$$y = g(X) + e$$

where $g(x)$ is the conditional mean function and e is the regression error. Let $f_e(e \mid x)$ be the conditional density of e given $X = x$. Then the conditional density of y is

$$f(y \mid x) = f_e(y - g(x) \mid x)$$

That is, we can write the conditional density of y in terms of the regression function and the conditional density of the error.

This decomposition suggests an alternative two-step estimator of f. First, estimate g. Second, estimate f_e.

The estimator $\hat{g}(x)$ for g can be NW, WNW, or LL.

The residuals are $\hat{e}_i = y_i - \hat{g}(X_i)$.

The second step is a conditional density estimator (NW, WNW or LL) applied to the residuals $\hat{e}_i$ as if they are observed data. This gives an estimator $\hat{f}_e(e \mid x)$.

The estimator for f is then

$$\hat{f}(y \mid x) = \hat{f}_e(y - \hat{g}(x) \mid x)$$

The first-order asymptotic distribution of $\hat{f}$ turns out to be identical to the ideal case where e_i is directly observed. This is because the first step conditional mean estimator $\hat{g}(x)$ converges at a rate faster than the second step estimator (at least if the first step is done with a bandwidth of the optimal order). e.g. if $q = 1$ then $\hat{g}(x)$ is optimally computed with a bandwidth $h \sim n^{-1/5}$, so that $\hat{g}$ converges at the rate $n^{-2/5}$, yet the estimator $\hat{f}_e$ converges at the best rate $n^{-1/3}$, so the error induced by estimation of $\hat{g}$ is of lower stochastic order.

The gain from the two-step estimator is that the conditional density of e typically has less dependence on X than the conditional density of y itself. This is because the conditional mean $g(X)$ has been removed, leaving only the higher-order dependence. The accuracy of nonparametric estimation improves as the estimated function becomes smoother and less dependent on the conditioning variables. Partially this occurs because reduced dependence allows for larger bandwidths, which reduces estimation variance.)

As an extreme case, if $f_e(e \mid x) = f_e(e \mid x)$ does not depend on one of the X variables, the $\hat{f}_e$ can converge at the $n^{-2/(q+4)}$ rate of the conditional mean. In this case the two-step estimator actually has an improved rate of convergence relative to the conventional estimator.

Two-step estimators of this form are often employed in practical applications, but do not seem to have been discussed much in the theoretical literature.

We could also consider a 3-step estimator, based on the expressions

$$\begin{aligned} y &= g(X) + e \\ e^2 &= \sigma^2(X) + \eta \\ \eta &\mid x \sim f_\eta(\eta \mid x) \end{aligned}$$

$$f(y \mid x) = f_\eta\left(\frac{y - g(x)}{\sigma(x)} \mid x\right)$$

The 3-step estimator is: First $\hat{g}(x)$ by nonparametric regression, obtain residuals $\hat{e}_i$. Second, $\hat{\sigma}^2(x)$ by nonparametric regression using $\hat{e}_i^2$ as dependent variable. Obtain rescaled residuals $\hat{\eta}_i = \hat{e}_i / \hat{\sigma}(X_i)$. Third, $\hat{f}_\eta(\eta \mid x)$ as the nonparametric conditional density estimator for $\hat{\eta}_i$. Then we can set

$$\hat{f}(y \mid x) = \hat{f}_\eta\left(\frac{y - \hat{g}(x)}{\hat{\sigma}(x)} \mid x\right)$$

In cases of strong variance effects (such as in financial data) this method may be desireable.

As the variance estimator $\hat{\sigma}^2(x)$ converges at the same rate as the mean $\hat{g}(x)$, the same first-order properties apply to the 3-step estimator as to the 2-step estimator. Namely, f_η should have reduced dependence on x, so it should be relatively well estimated even with large x-bandwidths, resulting in reduced MSE relative to the 1-step and 2-step estimators.

Given these insights, it might seem sensible to apply the 2-step or 3-step idea to conditional distribution estimation. Unfortunately the analysis is not quite as simple. In this setting, the nonparametric conditional mean, conditional variance, and conditional distribution estimators all converge at the same rates. Thus the distribution of the estimate of the CDF of e_i depends on the fact that it is a 2-step estimator, and it is not immediately obvious how this affects the asymptotic distribution. I have not seen an investigation of this issue.

6 Conditional Quantile Estimation

6.1 Quantiles

Suppose Y is univariate with distribution F.

If F is continuous and strictly increasing then its inverse function is uniquely defined. In this case the α'th quantile of Y is

$$q_\alpha = F^{-1}(\alpha).$$

If F is not strictly increasing then the inverse function is not well defined and thus quantiles are not unique but are interval-valued. To allow for this case it is conventional to simply define the quantile as the lower bound of this endpoint. Thus the general definition of the α'th quantile is

$$q_\alpha = \inf \left\{ y : F(y) \geq \alpha \right\}.$$

Quantiles are functions from probabilities to the sample space, and monotonically increasing in α.

Multivariate quantiles are not well defined. Thus quantiles are used for univariate and conditional settings.

If you know a distribution function F then you know the quantile function q_α. If you have an estimate $\hat{F}(y)$ of $F(y)$ then you can define the estimate

$$\hat{q}_\alpha = \inf \left\{ y : \hat{F}(y) \geq \alpha \right\}$$

If $\hat{F}(y)$ is monotonic in y then $\hat{q}_\alpha$ will also be monotonic in α. When a smoothed estimator $\hat{F}(y)$ is used, then we can write the quantile estimator more simply as $\hat{q}_\alpha = \hat{F}^{-1}(\alpha)$.

Suppose that $\hat{F}(y)$ is the (unsmoothed) EDF from a sample of size n. In this case, $\hat{q}_\alpha$ equals $Y_{([n\alpha])}$, the $[\alpha n]$'th order statistic from the sample. If αn is not an integer, $[\alpha n]$ is the greatest integer less than αn. We could also view the interval $\left[Y_{([n\alpha])}, Y_{([n\alpha]+1)} \right]$ as the quantile estimate. We ignore these distinctions in practice.

When $\hat{F}(y)$ is the EDF we can also write the quantile estimator as

$$\hat{q}_\alpha = \underset{q}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\alpha \left(Y_i - q \right)$$

where

$$\begin{aligned} \rho_\alpha(u) &= u \left[\alpha - 1 \left(u \leq 0 \right) \right] \\ &= \begin{cases} -\left(1-\alpha\right) u & u < 0 \\ \\ \alpha u & u \geq 0 \end{cases} \end{aligned}$$

is called the “check function”.

6.2 Conditional Quantiles

If the conditional distribution of Y given $X = x$ is $F(y \mid x)$ then the conditional quantile of Y given $X = x$ is

$$\begin{aligned} q_\alpha &= \inf \{y : F(y \mid x) \geq \alpha\} \\ &= F^{-1}(\alpha \mid x) \end{aligned}$$

Conditional quantiles are functions from probabilities to the sample space, for a fixed value of the conditioning variables.

One method for nonparametric conditional quantile estimation is to invert an estimated distribution function. Take an estimate $\hat{F}(y \mid x)$ of $F(y \mid x)$. Then we can define

$$\hat{q}_\alpha(x) = \inf \left\{ y : \hat{F}(y \mid x) \geq \alpha \right\}$$

When $\hat{F}(y \mid x)$ is smooth in y we can write this as $\hat{q}_\alpha(x) = \hat{F}^{-1}(\alpha \mid x)$.

This method is particularly appropriate for inversion of the smoothed CDF estimators $\tilde{F}(y \mid x)$.

This inversion method requires that $\hat{F}(y \mid x)$ be a distribution function (that it lies in $[0, 1]$ and is monotonic), which is not ensured if $\hat{F}(y \mid x)$ is computed by LL. The NW, WNW and smoothed versions are all appropriate. When $\hat{F}(y \mid x)$ is a distribution function then $\hat{q}_\alpha(x)$ will satisfy the properties of a quantile function.

6.3 Check Function Approach

Another estimation method is to define a weighted check function. This can be done using either a locally constant or locally linear specification.

The locally constant (NW) method uses the criterion

$$S_\alpha(q \mid x) = \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) \rho_\alpha(Y_i - q)$$

It is a locally weighted the check function, for observations “close to” $X_i = x$. The nonparametric quantile estimator is

$$\hat{q}_\alpha(x) = \underset{q}{\operatorname{argmin}}\, S_\alpha(q \mid x).$$

The local linear (LL) criterion is

$$S_\alpha(q, \beta \mid x) = \sum_{i=1}^{n} K\left(H^{-1}(X_i - x)\right) \rho_\alpha\left(Y_i - q - (X_i - x)' \beta\right).$$

The estimator is

$$\left\{\hat{q}_\alpha\left(x\right),\hat{\beta}_\alpha(x)\right\}=\underset{q,\beta}{\operatorname{argmin}}\,S_\alpha\left(q,\beta\mid x\right).$$

The conditional quantile estimator is $\hat{q}_\alpha\left(x\right)$, with derivative estimate $\hat{\beta}_\alpha(x)$. Numerically, these problems are identical to weighted linear quantile regression.

6.4 Asymptotic Distribution

The asymptotic distributions of the quantile estimators are scaled versions of the asymptotic distributions of the CDF estimators (see the Li-Racine text for details).

The CDF inversion method and the check function method have the same asymptotic distributions.

The asymptotic bias of the quantile estimators depends on whether a local constant or local linear method was used, and whether smoothing in the y direction is used.

6.5 Bandwidth Selection

Optimal bandwidth selection for nonparametric quantile regression is less well studied than the other methods.

As the asymptotic distributions seem to be scaled versions of the CDF estimators, and the quantile estimator can be viewed as a by-product of CDF estimation, it seems reasonable to select bandwidths by a method optimal for CDF estimation, e.g. cross-validation for conditional distribution function estimation.

7 Semiparametric Methods and Partially Linear Regression

7.1 Overview

A model is called semiparametric if it is described by θ and τ where θ is finite-dimensional (e.g. parametric) and τ is infinite-dimensional (nonparametric). All moment condition models are semiparametric in the sense that the distribution of the data (τ) is unspecified and infinite dimensional. But the settings more typically called "semiparametric" are those where there is explicit estimation of τ.

In many contexts the nonparametric part τ is a conditional mean, variance, density or distribution function.

Often θ is the parameter of interest, and τ is a nuisance parameter, but this is not necessarily the case.

In many semiparametric contexts, τ is estimated first, and then $\hat{\theta}$ is a two-step estimator. But in other contexts (θ, τ) are jointly estimated.

7.2 Feasible Nonparametric GLS

A classic semiparametric model, which is not in Li-Racine, is feasible GLS with unknown variance function. The seminal papers are Carroll (1982, Annals of Statistics) and Robinson (1987, Econometrica). The setting is a linear regression

$$\begin{aligned} y_i &= X_i'\theta + e_i \\ \mathrm{E}\left(e_i \mid X_i\right) &= 0 \\ \mathrm{E}\left(e_i^2 \mid X_i\right) &= \sigma^2\left(X_i\right) \end{aligned}$$

where the variance function $\sigma^2(x)$ is unknown but smooth in x, and $x \in \mathbb{R}^q$. (The idea also applies to non-linear but parametric regression functions). In this model, the nonparametric nuisance parameter is $\tau = \sigma^2(\cdot)$

As the model is a regression, the efficiency bound for θ is attained by GLS regression

$$\tilde{\theta} = \left(\sum_{i=1}^{n} \frac{1}{\sigma^2\left(X_i\right)} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} \frac{1}{\sigma^2\left(X_i\right)} X_i y_i\right).$$

This of course is infeasible. Carroll and Robinson suggested replacing $\sigma^2(X_i)$ with $\hat{\sigma}^2(X_i)$ where $\hat{\sigma}^2(x)$ is a nonparametric estimator. (Carroll used kernel methods; Robinson used nearest neighbor methods.) Specifically, letting $\hat{\sigma}^2(x)$ be the NW estimator of $\sigma^2(x)$, we can define the feasible estimator

$$\hat{\theta} = \left(\sum_{i=1}^{n} \frac{1}{\hat{\sigma}^2\left(X_i\right)} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} \frac{1}{\hat{\sigma}^2\left(X_i\right)} X_i y_i\right).$$

This seems sensible. The question is find its asymptotic distribution, and in particular to find if it is asymptotically equivalent to $\tilde{\theta}$.

7.3 Generated Regressors

The model is

$$\begin{aligned} y_i &= \theta\tau(X_i) + e_i \\ \mathrm{E}\left(e_i \mid X_i\right) &= 0 \end{aligned}$$

where θ is finite dimensional but τ is an unknown function. Suppose τ is identified by another equation so that we have a consistent estimate of $\hat{\tau}(x)$ for $\tau(x)$.(Imagine a non-parametric Heckman estimator).

Then we could estimate θ by least-squares of y_i on $\hat{\tau}(Z_i)$.

This problem is called generated regressors, as the regressor is a (consistent) estimate of a infeasible regressor.

In general, $\hat{\theta}$ is consistent. But what is its distribution?

7.4 Andrews' MINPIN Theory

A useful framework to study the type of problem from the previous section is given in Andrews (Econometrica, 1994). It is reviewed in section 7.3 of Li-Racine but the discussion is incomplete and there is at least one important omission. If you really want to learn this theory I suggest reading Andrews' paper.

The setting is when the estimator $\hat{\theta}$ MINimizes a criterion function which depends on a Preliminary Infinite dimensional Nuisance parameter estimator, hence MINPIN.

Let $\theta \in \Theta$ be the parameter of interest, let $\tau \in T$ denote the infinite-dimensional nuisance parameter. Let θ_0 and τ_0 denote the true values.

Let $\hat{\tau}$ be a first-step estimate of τ, and assume that it is consistent: $\hat{\tau} \rightarrow_p \tau_0$

Now suppose that the criterion function for estimation of θ depends on the first-step estimate $\hat{\tau}$. Let the criterion be $Q_n\left(\theta, \tau\right)$ and suppose that

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, Q_n\left(\theta, \hat{\tau}\right)$$

Thus $\hat{\theta}$ is a two-step estimator.

Assume that

$$\begin{aligned} \frac{\partial}{\partial\theta} Q_n\left(\theta, \tau\right) &= \bar{m}_n\left(\theta, \tau\right) + o_p\left(\frac{1}{\sqrt{n}}\right) \\ \bar{m}_n\left(\theta, \tau\right) &= \frac{1}{n}\sum_{i=1}^{n} m_i\left(\theta, \tau\right) \end{aligned}$$

where $m_i(\theta, \tau)$ is a function of the i'th observation. In just-identified models, there is no $o_p\left(\frac{1}{\sqrt{n}}\right)$ error (and we now ignore the presence of this error).

In the FGLS example, $\tau(\cdot) = \sigma^2(\cdot)$, $\hat{\tau}(\cdot) = \hat{\sigma}^2(\cdot)$, and

$$m_i(\theta, \tau) = \frac{1}{\tau(X_i)} X_i \left(y_i - X_i'\theta\right).$$

In the generated regressor problem,

$$m_i(\theta, \tau) = \tau(X_i)\left(y_i - \theta\tau(X_i)\right).$$

In general, the first-order condition (FOC) for $\hat{\theta}$ is

$$0 = \bar{m}_n\left(\hat{\theta}, \hat{\tau}\right)$$

Assume $\hat{\theta} \rightarrow_p \theta_0$. Implicit to obtain consistency is the requirement that the population expecation of the FOC is zero, namely

$$\mathrm{E}m_i(\theta_0, \tau_0) = 0$$

and we assume that this is the case.

We expand the FOC in the first argument

$$\begin{aligned} 0 &= \sqrt{n}\bar{m}_n\left(\hat{\theta}, \hat{\tau}\right) \\ &= \sqrt{n}\bar{m}_n(\theta_0, \hat{\tau}) + M_n(\theta_0, \hat{\tau})\sqrt{n}\left(\hat{\theta} - \theta_0\right) + o_p(1) \end{aligned}$$

where

$$M_n(\theta, \tau) = \frac{\partial}{\partial \theta'}\bar{m}_n(\theta, \tau)$$

It follows that

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \simeq -M_n(\theta_0, \hat{\tau})^{-1}\sqrt{n}\bar{m}_n(\theta_0, \hat{\tau}).$$

If $M_n(\theta, \tau)$ converges to its expectation

$$M(\theta, \tau) = \mathrm{E}\frac{\partial}{\partial \theta'} m_i(\theta, \tau)$$

uniformly in its arguments, then $M_n(\theta_0, \hat{\tau}) \rightarrow_p M(\theta_0, \tau_0) = M$, say. Then

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \simeq -M^{-1}\sqrt{n}\bar{m}_n(\theta_0, \hat{\tau}).$$

We cannot take a Taylor expansion in τ because it is infinite dimensional. Instead, Andrews uses a stochastic equicontinuity argument. Define the population version of $\bar{m}_n(\theta, \tau)$

$$m(\theta, \tau) = \mathrm{E}m_i(\theta, \tau).$$

Note that at the true values $m(\theta_0, \tau_0) = 0$ (as discussed above) but the function is non-zero for generic values.

Define the function

$$\begin{aligned} \nu_n(\tau) &= \sqrt{n}\left(\bar{m}_n(\theta_0, \tau) - m(\theta_0, \tau)\right) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(m_i(\theta_0, \tau) - \mathrm{E} m_i(\theta_0, \tau)\right) \end{aligned}$$

Notice that $\nu_n(\tau)$ is a normalized sum of mean-zero random variables. Thus for any fixed τ, $\nu_n(\tau)$ converges to a normal random vector. Viewed as a function of τ, we might expect $\nu_n(\tau)$ to vary smoothly in the argument τ. The stochastic formulation of this is called stochastic equicontinuity. Roughly, as $n \to \infty$, $\nu_n(\tau)$ remains well-behaved as a function of τ. Andrews first key assumption is that $\nu_n(\tau)$ is stochastically equicontinuous.

The important implication of stochastic equicontinuity is that $\hat{\tau} \to_p \tau_0$ implies

$$\nu_n(\hat{\tau}) - \nu_n(\tau_0) \to_p 0.$$

Intuitively, if $g(\tau)$ is continuous, then $g(\hat{\tau}) - g(\tau_0) \to_p 0$. More generally, if $g_n(\tau)$ converges in probability uniformly to a continuous function $g(\tau)$ then $g_n(\hat{\tau}) - g(\tau_0) \to_p 0$. The case of stochastic equicontinuity is the most general, but still has the same implication.

Since $\mathrm{E} m_i(\theta_0, \tau_0) = 0$, when we evaluate the empirical process at the true value τ_0 we have a zero-mean normalized sum, which is asymptotically normal by the CLT:

$$\begin{aligned} \nu_n(\tau_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(m_i(\theta_0, \tau_0) - \mathrm{E} m_i(\theta_0, \tau_0)\right) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_i(\theta_0, \tau_0) \\ &\to_d N(0, \Omega) \end{aligned}$$

where

$$\Omega = \mathrm{E} m_i(\theta_0, \tau_0)\, m_i(\theta_0, \tau_0)'$$

It follows that

$$\nu_n(\hat{\tau}) = \nu_n(\tau_0) + o_p(1) \to_d N(0, \Omega)$$

and thus

$$\begin{aligned} \sqrt{n}\bar{m}_n(\theta_0, \hat{\tau}) &= \sqrt{n}\left(\bar{m}_n(\theta_0, \hat{\tau}) - m(\theta_0, \hat{\tau})\right) + \sqrt{n} m(\theta_0, \hat{\tau}) \\ &= \nu_n(\hat{\tau}) + \sqrt{n} m(\theta_0, \hat{\tau}) \end{aligned}$$

The final detail is what to do with $\sqrt{n}m\left(\theta_0,\hat{\tau}\right)$. Andrews directly assumes that

$$\sqrt{n}m\left(\theta_0,\hat{\tau}\right)\rightarrow_p 0$$

This is the second key assumption, and we discuss it below. Under this assumption,

$$\sqrt{n}\bar{m}_n\left(\theta_0,\hat{\tau}\right)\rightarrow_d N\left(0,\Omega\right)$$

and combining this with our earlier expansion, we obtain:

Andrews MINPIN Theorem. Under Assumptions 1-6 below,

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)\rightarrow_p N\left(0,V\right)$$

where

$$\begin{aligned}V &= M^{-1}\Omega M^{-1\prime}\\ M &= \mathrm{E}\frac{\partial}{\partial\theta'}m_i\left(\theta_0,\tau_0\right)\\ \Omega &= \mathrm{E}m_i\left(\theta_0,\tau_0\right)m_i\left(\theta_0,\tau_0\right)'\end{aligned}$$

Assumption 1 *As $n\rightarrow\infty$, for $m\left(\theta,\tau\right)=\mathrm{E}m_i\left(\theta,\tau\right)$*

1. $\hat{\theta}\rightarrow_p\theta_0$
2. $\hat{\tau}\rightarrow_p\tau_0$
3. $\sqrt{n}m\left(\theta_0,\hat{\tau}\right)\rightarrow_p 0$
4. $m\left(\theta_0,\tau_0\right)=0$
5. *$\nu_n\left(\tau\right)$ is stochastically equicontinuous at τ_0.*
6. *$\bar{m}_n\left(\theta,\tau\right)$ and $\frac{\partial}{\partial\theta}\bar{m}_n\left(\theta,\tau\right)$ satisfy uniform WLLN's over $\Theta\times T$. (They converge in probability to their expectations, uniformly over the parameter space.)*

Discussion of result: The theorem says that the semiparametric estimator $\hat{\theta}$ has the same asymptotic distribution as the idealized estimator obtained by replacing the nonparametric estimate $\hat{\tau}$ with the true function τ_0. Thus the estimator is adaptive. This might seem too good to be true. The key is assumption, which holds in some cases, but not in others.

Discussion of assumptions.

Assumptions 1 and 2 state that the estimators are consistent, which should be separately verified. Assumption 4 states that the FOC identifies θ when evaluated at the true τ_0. Assumptions 5 and 6 are regularity conditions, essentially smoothness of the underlying functions, plus sufficient moments.

7.5 Orthogonality Assumption

The key assumption 3 for Andrews' MINPIN theorey was somehow missed in the write-up in Li-Racine. Assumption 3 is not always true, and is not just a regularity condition. It requires a sort of orthogonality between the estimation of θ and τ.

Suppose that τ is finite-dimensional. Then by a Taylor expansion

$$\begin{aligned}\sqrt{n}m\left(\theta_0,\hat{\tau}\right) &\simeq \sqrt{n}m\left(\theta_0,\tau_0\right)+\frac{\partial}{\partial\tau}m\left(\theta_0,\tau\right)'\sqrt{n}\left(\hat{\tau}-\tau_0\right)\\ &= \frac{\partial}{\partial\tau}m\left(\theta_0,\tau_0\right)'\sqrt{n}\left(\hat{\tau}-\tau_0\right)\end{aligned}$$

since $m\left(\theta_0,\tau_0\right)=0$ by Assumption 4. Since τ is parametric, we should expect $\sqrt{n}\left(\hat{\tau}-\tau_0\right)$ to converge to a normal vector. Thus this expression will converge in probability to zero only if

$$\frac{\partial}{\partial\tau}m\left(\theta_0,\tau_0\right)=0$$

Recall that $m\left(\theta,\tau\right)$ is the expectation of the FOC, which is the derivative of the criterion wrt θ. Thus $\frac{\partial}{\partial\tau}m\left(\theta,\tau\right)$ is the cross-derivative of the criterion, and the above statement is that this cross-derivative is zero, which is an orthogonality condition (e.g. block diagonality of the Hessian.)

Now when τ is infinite-dimensional, the above argument does not work, but it lends intuition.

An analog of the derivative condition, which is sufficient for Assumption 3, is that

$$m\left(\theta_0,\tau\right)=0$$

for all τ in a neighborhood of τ_0.

In the FGLS example,

$$m\left(\theta_0,\tau\right)=\mathrm{E}\left(\frac{1}{\tau\left(X_i\right)}X_ie_i\right)=0$$

for all τ, so this is Assumption 3 is satisfied in this example. We have the implication:

Theorem. Under regularity conditions the Feasible nonparametric GLS estimator of the previous section is asymptotically equivalent to infeasible GLS.

In the generated regressor example

$$\begin{aligned}m\left(\theta_0,\tau\right) &= \mathrm{E}\left(\tau(X_i)\left(y_i-\theta_0\tau(X_i)\right)\right)\\ &= \mathrm{E}\left(\tau(X_i)\left(e_i+\theta_0\left(\tau_0(X_i)-\tau(X_i)\right)\right)\right)\\ &= \theta_0\mathrm{E}\left(\tau(X_i)\left(\tau_0(X_i)-\tau(X_i)\right)\right)\\ &= \theta_0\int\tau(x)\left(\tau_0(x)-\tau(x)\right)f(x)dx\end{aligned}$$

Assumption 3 requires $\sqrt{n}m\left(\theta_0,\hat{\tau}\right)\rightarrow_p 0$. But note that $\sqrt{n}m\left(\theta_0,\hat{\tau}\right)\simeq\sqrt{n}\left(\tau_0(x)-\hat{\tau}(x)\right)$ which certainly does not converge to zero. Assumption 3 is generically violated when there are generated regressors.

There is one interesting exception. When $\theta_0 = 0$ then $m(\theta_0, \tau) = 0$ and thus $\sqrt{n}m(\theta_0, \hat{\tau}) = 0$ so Assumption 3 is satisified.

We see that Andrews' MINPIN assumption do not apply in all semiparametric models. Only those which satisfy Assumption 3, which needs to be verified. The other key condition is stochastic equicontinuity, which is difficult to verify but is generaly satisfied for "well-behaved" estimators. The remaining assumptions are smoothness and regularity conditions, and typically are not of concern in applications.

7.6 Partially Linear Regression Model

The semiparametric partially linear regression model is

$$\begin{aligned} y_i &= X_i'\beta + g(Z_i) + e_i \\ \mathrm{E}(e_i \mid X_i, Z_i) &= 0 \\ \mathrm{E}\left(e_i^2 \mid X_i = x, Z_i = z\right) &= \sigma^2(x, z) \end{aligned}$$

That is, the regressors are (X_i, Z_i), and the model specifies the conditional mean as linear in X_i but possibly non-linear in $Z_i \in \mathbb{R}^q$. This is a very useful compromise between fully nonparametric and fully parametric. Often the binary (dummy) variables are put in X_i. Often there is just one nonlinear variable: $q = 1$, to keep things simple.

The goal is to estimate β and g, and to obtain confidence intervals.

The first issue to consider is identification. Since g is unconstrained, the elements of X_i cannot be collinear with any function of Z_i. This means that we must exclude from X_i intercepts and any deterministic function of Z_i. The function g includes these components.

7.7 Robinson's Transformation

Robinson (Econometrica, 1988) is the seminal treatment of the partially linear model. He first contribution is to show that we can concentrate out the unknown g by using a genearlization of residual regression.

Take the equation

$$y_i = X_i'\beta + g(Z_i) + e_i$$

and apply the conditional expectation operator $\mathrm{E}(\cdot \mid Z_i)$. We obtain

$$\begin{aligned} \mathrm{E}(y_i \mid Z_i) &= \mathrm{E}\left(X_i'\beta \mid Z_i\right) + \mathrm{E}(g(Z_i) \mid Z_i) + \mathrm{E}(e_i \mid Z_i) \\ &= \mathrm{E}(X_i \mid Z_i)'\beta + g(Z_i) \end{aligned}$$

(using the law of iterated expectations). Defining the conditional expectations

$$\begin{aligned} g_y(z) &= \mathrm{E}(y_i \mid Z_i = z) \\ g_x(z) &= \mathrm{E}(X_i \mid Z_i = z) \end{aligned}$$

We can write this expression as

$$g_y(z) = g_x(z)' \beta + g(z)$$

Subtracting from the original equation, the function g disappears:

$$y_i - g_y(Z_i) = (X_i - g_x(Z_i))' \beta + e_i$$

We can write this as

$$\begin{aligned} e_{yi} &= e_{xi}' \beta + e_i \\ y_i &= g_y(Z_i) + e_{yi} \\ X_i &= g_x(Z_i) + e_{xi} \end{aligned}$$

That is, β is the coefficient of the regression of e_{yi} on e_{xi}, where these are the conditional expectations errors from the regression of y_i (and X_i) on Z_i only.

This is a conditional expecation generalization of the idea of residual regression.

This transformed equation immediately suggests an infeasible estimator for β, by LS of e_{yi} on e_{xi} :

$$\tilde{\beta} = \left(\sum_{i=1}^n e_{xi} e_{xi}' \right)^{-1} \sum_{i=1}^n e_{xi} e_{yi}$$

7.8 Robinson's Estimator

Robinson suggested first estimating g_y and g_x by NW regression, using these to obtain residuals $\hat{e}_{xi}$ and $\hat{e}_{xi}$, and replacing these in the formula for $\tilde{\beta}$.

Specifically, let $\hat{g}_y(z)$ and $\hat{g}_x(z)$ denote the NW estimates of the conditional mean of y_i and X_i given $Z_i = z$. Assuming $q = 1$,

$$\hat{g}_y(z) = \frac{\sum_{i=1}^n k\left(\frac{Z_i - z}{h} \right) y_i}{\sum_{i=1}^n k\left(\frac{Z_i - z}{h} \right)}$$

$$\hat{g}_x(z) = \frac{\sum_{i=1}^n k\left(\frac{Z_i - z}{h} \right) X_i}{\sum_{i=1}^n k\left(\frac{Z_i - z}{h} \right)}$$

The estimator $\hat{g}_x(z)$ is a vector of the same dimension as X_i.

Notice that you are regressing each variable (the y_i and the X_i's) separately on the continuous variable Z_i. You should view each of these regressions as a separate NW regression. You should probably use a different bandwidth h for each of these regressions, as the depdendence on Z_i will depend on the variable. (For example, some regressors X_i might be independent of Z_i so you would want to use an infinite bandwidth for those cases.) While Robinson discussed NW regression, this is not essential. You could substitute LL or WNW instead.

Given the regression functions, we obtain the regression residuals

$$\begin{aligned}\hat{e}_{yi} &= y_i - \hat{g}_y(Z_i) \\ \hat{e}_{xi} &= X_i - \hat{g}_x(Z_i)\end{aligned}$$

Our first attempt at an estimator for β is then

$$\hat{\beta} = \left(\sum_{i=1}^n \hat{e}_{xi}\hat{e}_{xi}'\right)^{-1} \sum_{i=1}^n \hat{e}_{xi}\hat{e}_{yi}$$

7.9 Trimming

The asymptotic theory for semiparametric estimators, typically requires that the first step estimator converges uniformly at some rate. The difficulty is that $\hat{g}_y(z)$ and $\hat{g}_x(z)$ do not converge uniformly over unbounded sets. Equivalently, the problem is due to the estimated density of Z_i in the denominator. Another way of viewing the problem is that these estimates are quite noisy in sparse regions of the sample space, so residuals in such regions are noisy, and this could unduely influence the estimate of β.

suffer a problem that they can be unduly influence by unstable residuals from observations in sparse regions of the sample space. The nonparametric regression estimates depend inversely on the density estimate

$$\hat{f}_z(z) = \frac{1}{nh}\sum_{i=1}^n k\left(\frac{Z_i - z}{h}\right).$$

For values of z where $f_z(z)$ is close to zero, $\hat{f}_z(z)$ is not bounded away from zero, so the NW estimates at this point can be poor. Consequently the residuals for observations i such that $f_z(Z_i)$ will be quite unreliable, and can have an undue influence on $\hat{\beta}$.

A standard solution is to introduce "trimming". Let

$$\hat{f}_z(z) = \frac{1}{nh}\sum_{i=1}^n k\left(\frac{Z_i - z}{h}\right),$$

let $b > 0$ be a trimming constant and let $1_i(b) = 1\left(\hat{f}_z(Z_i) \geq b\right)$ denote a indicator variable for those observations for which the estimated density of Z_i is above b.

The trimmed version of $\hat{\beta}$ is

$$\hat{\beta} = \left(\sum_{i=1}^{n} \hat{e}_{xi} \hat{e}_{xi}' 1_i(b) \right)^{-1} \sum_{i=1}^{n} \hat{e}_{xi} \hat{e}_{yi} 1_i(b)$$

This is a trimmed LS residual regression.

The asymptotic theory requires that $b = b_n \to 0$ but unfortunately there is not good guidance about how to select b in practice. Often trimming is ignored in applications. One practical suggestion is to estimate β with and without trimming to assess robustness.

7.10 Asymptotic Distribution

Robinson (1988), Andrews (1994) and Li (1996) are references. The needed regularity conditions are that the data are iid, Z_i has a density, and the regression functions, density, and conditional variance function are sufficiently smooth with respect to their arguments. Assuming a second-order kernel, and for simplicity writing $h = h_1 = \cdots = h_q$, the important condition on the bandwidths sequence is

$$\sqrt{n} \left(h^4 + \frac{1}{nh^q} \right) \to 0$$

Technically, this is not quite enough, as this ignores the interaction with the trimming parameter b. But since this can be set $b_n \to 0$ at an extremely slow rate, it can be safely ignored. The above condition is similar to the standard convergence rates for nonparametric estimation, multiplied by $\sqrt{n}$. Equivalently, what is essential is that the uniform MSE of the nonparametric estimators converge faster than $\sqrt{n}$, or that the estimators themselves converges faster than $n^{-1/4}$. That is, what we need is

$$n^{-1/4} \sup_z |\hat{g}_y(z) - g_y(z)| \to_p 0$$

$$n^{-1/4} \sup_z |\hat{g}_x(z) - g_x(z)| \to_p 0$$

From the theory for nonparametric regression, these rates hold when bandwidths are picked optimally and $q \leq 3$.

In practice, $q \leq 3$ is probably sufficient. If $q > 3$ is desired, then higher-order kernels can be used to improve the rate of convergence. So long as the rate is faster than $n^{-1/4}$, the following result applies.

Theorem (Robinson). Under regularity conditions, including $q \leq 3$, the trimmed estimator satisfies

$$\sqrt{n} \left(\hat{\beta} - \beta \right) \to_d N(0, V)$$

$$V = \left(E\left(e_{xi} e_{xi}' \right) \right)^{-1} \left(E\left(e_{xi} e_{xi}' \sigma^2 (X_i, Z_i) \right) \right)^{-1} \left(E\left(e_{xi} e_{xi}' \right) \right)^{-1}$$

That is, $\hat{\beta}$ is asymptotically equivalent to the infeasible estimator $\tilde{\beta}$.

The variance matrix may be estimated using conventional LS methods.

7.11 Verification of Andrews' MINPIN Condition

This Theorem states that Robinson's two-step estimator for β is asymptotically equivalent to the infeasible one-step estimator. This is an example of the application of Andrews' MINPIN theory. Andrews specifically mentions that the $n^{-1/4}$ convergence rates for $\hat{g}_y(z)$ *and* $\hat{g}_x(z)$ are essential to obtain this result.

To see this, note that the estimator $\hat{\beta}$ solves the FOC

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\hat{g}_x(Z_i)\right)\left(y_i-\hat{g}_y(Z_i)-\hat{\beta}'\left(X_i-\hat{g}_x(Z_i)\right)\right)=0$$

In Andrews MINPIN notation, let $\tau_x=\hat{g}_x$ and $\tau_y=\hat{g}_y$ denote fixed (function) values of the regression estimates, then

$$m_i(\theta_0,\tau)=\left(X_i-\tau_x(Z_i)\right)\left(y_i-\tau_y(Z_i)-\theta_0'\left(X_i-\tau_x(Z_i)\right)\right)$$

Since

$$y_i=g_y(Z_i)+\left(X_i-g_x(Z_i)\right)'\beta+e_i$$

then

$$\mathrm{E}\left(y_i-\tau_y(Z_i)-\theta_0'\left(X_i-\tau_x(Z_i)\right)\mid X_i,Z_i\right)=g_y(Z_i)-\tau_y(Z_i)-\left(g_x(Z_i)-\tau_x(Z_i)+\right)'\theta_0$$

and

$$\begin{aligned}
m(\theta_0,\tau) &= \mathrm{E}m_i(\theta_0,\tau)\\
&= \mathrm{EE}\left(m_i(\theta_0,\tau)\mid X_i,Z_i\right)\\
&= \mathrm{E}\left(\left(X_i-\tau_x(Z_i)\right)\left(g_y(Z_i)-\tau_y(Z_i)-\left(g_x(Z_i)-\tau_x(Z_i)\right)'\theta_0\right)\right)\\
&= \mathrm{E}\left(\left(g_x(Z_i)-\tau_x(Z_i)\right)\left(g_y(Z_i)-\tau_y(Z_i)-\left(g_x(Z_i)-\tau_x(Z_i)\right)'\theta_0\right)\right)\\
&= \int\left(\left(g_x(z)-\tau_x(z)\right)\left(g_y(z)-\tau_y(z)-\left(g_x(z)-\tau_x(z)\right)'\theta_0\right)\right)f_z(z)dz
\end{aligned}$$

where the second-to-last line uses conditional expecations given X_i. Then replacing τ_x with $\hat{g}_x$ and τ_y with $\hat{g}_y$

$$\sqrt{n}m(\theta_0,\hat{\tau})=\int\left(\left(g_x(z)-\hat{g}_x(z)\right)\left(g_y(z)-\hat{g}_y(z)-\left(g_x(z)-\hat{g}_x(z)\right)'\theta_0\right)\right)f_z(z)dz$$

Taking bounds

$$\left|\sqrt{n}m(\theta_0,\hat{\tau})\right|\leq\left(\sup_z\left|g_x(z)-\hat{g}_x(z)\right|\sup_z\left|g_y(z)-\hat{g}_y(z)\right|+\sup_z\left|g_x(z)-\hat{g}_x(z)\right|^2\right)\sup_z\left|f_y(z)\right|\to_p 0$$

when the nonparametric regression estimates converge faster than $n^{-1/4}$.

Indeed, we see that the $o(n^{-1/4})$ convergence rates imply the key condition $\sqrt{n}m\left(\theta_0, \hat{\tau}\right) \to_p 0$.

7.12 Estimation of Nonparametric Component

Recall that the model is

$$y_i = X_i'\beta + g\left(Z_i\right) + e_i$$

and the goal is to estimate β and g. We have described Robinson's estimator for β. We now discuss estimation of g.

Since $\hat{\beta}$ converges at rate $n^{-1/2}$ which is faster than a nonparametric rate, we can simply pretend that β is known, and do nonparametric regression of $y_i - X_i'\hat{\beta}$ on $Z_i - z$

$$\hat{g}\left(z\right) = \frac{\sum_{i=1}^n k\left(\dfrac{Z_i - z}{h}\right)\left(y_i - X_i'\hat{\beta}\right)}{\sum_{i=1}^n k\left(\dfrac{Z_i - z}{h}\right)}$$

The bandwidth $h = (h_1, ..., h_q)$ is distinct from those for the first-stage regressions.

It is not hard to see that this estimator is asymptotically equivalent to the infeasible regressor when $\hat{\beta}$ is replaced with the true β_0.

Standard errors for $\hat{g}(x)$ may be computed as for standard nonparametric regression.

7.13 Bandwidth Selection

In a semiparametric context, it is important to study the effect a bandwidth has on the performance of the estimator of interest before determining the bandwidth. In many cases, this requires a nonconventional bandwidth rate.

However, this problem does not occur in the partially linear model. The first-step bandwidths h used for $\hat{g}_y\left(z\right)$ and $\hat{g}_x\left(z\right)$ are inputs for calculation of $\hat{\beta}$. The goal is presumably accurate estimation of β. The bandwith h impacts the theory for $\hat{\beta}$ through the uniform convergence rates for $\hat{g}_y\left(z\right)$ and $\hat{g}_x\left(z\right)$ – suggesting that we use conventional bandwidth rules, e.g. cross-validation.

8 Semiparametric Single Index Models

8.1 Index Models

A object of interest such as the conditional density $f(y \mid x)$ or conditional mean $\mathrm{E}(y \mid x)$ is a single index model when it only depends on the vector x through a single linear combination $x'\beta$.

Most parametric models are single index, including Normal regression, Logit, Probit, Tobit, and Poisson regression.

In a semiparametric single index model, the object of interest depends on x through the function $g(x'\beta)$ where $\beta \in \mathbb{R}^k$ and $g : \mathbb{R} \to \mathbb{R}$ are unknown. g is sometimes called a *link function.* In single index models, there is only one nonparametric dimension. These methods fall in the class of dimension reduction techniques.

The semiparametric single index regression model is

$$\mathrm{E}(y \mid x) = g\left(x'\beta\right) \tag{1}$$

where g is an unknown link function.

The semiparametric single index binary choice model is

$$P(y = 1 \mid x) = \mathrm{E}(y \mid x) = g\left(x'\beta\right) \tag{2}$$

where g is an unknown distribution function. We use g (rather than, say, F) to emphasize the connection with the regression model.

In both contexts, the function g includes any location and level shift, so the vector X_i cannot include an intercept. The level of β is not identified, so some normalization criterion for β is needed. It is typically easier to impose this on β than on g. One approach is to set $\beta'\beta = 1$. A second approach is to set one component of β to equal one. (This second approach requires that this variable correctly has a non-zero coefficient.)

The vector X_i must be dimension 2 or larger. If X_i is one-dimensional, then β is simply normalized to one, and the model is the one-dimensional nonparametric regression $\mathrm{E}(y \mid x) = g(x)$ with no semiparametric component.

Identification of β and g also requires that X_i contains at least one continuously distributed variable, and that this variable has a non-zero coefficient. If not, $X_i'\beta$ only takes a discrete set of values, and it would be impossible to identify a continuous function g on this discrete support.

8.2 Single Index Regression and Ichimura's Estimator

The semiparametric single index regression model is

$$\begin{aligned} y_i &= g\left(X_i'\beta\right) + e_i \\ \mathrm{E}(e_i \mid X_i) &= 0 \end{aligned}$$

This model generalizes the linear regression model (which sets $g(z)$ to be linear), and is a restriction of the nonparametric regression model.

The gain over full nonparametrics is that there is only one nonparametric dimension, so the curse of dimensionality is avoided.

Suppose g were known. Then you could estimate β by (nonlinear) least-squares. The LS criterion would be

$$S_n(\beta, g) = \sum_{i=1}^{n} \left(y_i - g\left(X_i'\beta\right)\right)^2.$$

We could think about replacing g with an estimate $\hat{g}$, but since $g(z)$ is the conditional mean of y_i given $X_i'\beta = z$, g depends on β, so a two-step estimator is likely to be inefficient.

In his PhD thesis, Ichimura proposed a semiparametric estimator, published later in the Journal of Econometrics (1993).

Ichimura suggested replacing g with the leave-one-out NW estimator

$$\hat{g}_{-i}\left(X_i'\beta\right) = \frac{\sum_{j\neq i} k\left(\frac{(X_j - X_i)'\beta}{h}\right) y_j}{\sum_{j\neq i} k\left(\frac{(X_j - X_i)'\beta}{h}\right)}.$$

The leave-one-out version is used since we are estimating the regression at the i'th observation i.

Since the NW estimator only converges uniformly over compact sets, Ichimura introduces trimming for the sum-of-squared errors. The criterion is then

$$S_n(\beta) = \sum_{i=1}^{n} \left(y_i - \hat{g}_{-i}\left(X_i'\beta\right)\right)^2 1_i(b)$$

He is not too specific about how to pick the trimming function, and it is likely that it is not important in applications.

The estimator of β is then

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, S_n(\beta).$$

The criterion is somewhat similar to cross-validation. Indeed, Hardle, Hall, and Ichimura (Annals of Statistics, 1993) suggest picking β and the bandwidth h jointly by minimization of $S_n(\beta)$.

In his paper, Ichimura claims that the $\hat{g}_{-i}\left(X_i'\beta\right)$ could be replaced by any other uniformly consistent estimator and the consistency of $\hat{\beta}$ would be maintained, but his asymptotic normality result would be lost. In particular, his proof rests on the asymptotic orthogonality of the derviative of $\hat{g}_{-i}\left(X_i'\beta\right)$ with e_i, which holds since the former is a leave-one-out estimator, and fails if it is a conventional NW estimator.

8.3 Asymptotic Distriubution of Ichimura's Estimator

Let β_0 denote the true value of β.

The tricky thing is that $\hat{g}_{-i}\left(X_i'\beta\right)$ is not estimating $g(X_i'\beta_0)$, rather it is estimating

$$G\left(X_i'\beta\right) = \mathrm{E}\left(y_i \mid X_i'\beta\right) = \mathrm{E}\left(g\left(X_i'\beta_0\right) \mid X_i'\beta\right)$$

the second equality since $y_i = g(X_i'\beta_0) + e_i$.

That is

$$G\left(z\right) = \mathrm{E}\left(y_i \mid X_i'\beta = z\right)$$

and $G\left(X_i'\beta\right)$ is then evaluated at $X_i'\beta$.

Note that

$$G\left(X_i'\beta_0\right) = g\left(X_i'\beta_0\right)$$

but for other values of β,

$$G\left(X_i'\beta\right) \neq g\left(X_i'\beta\right)$$

Hardle, Hall, and Ichimura (1993) show that the LS criterion is asymptotically equivalent to replacing $\hat{g}_{-i}\left(X_i'\beta\right)$ with $G\left(X_i'\beta\right)$, so

$$S_n\left(\beta\right) \simeq S_n^*\ \left(\beta\right) = \sum_{i=1}^{n}\left(y_i - G\left(X_i'\beta\right)\right)^2.$$

This approximation is essentially the same as Andrews' MINPIN argument, and relies on the estimator $\hat{g}_{-i}\left(X_i'\beta\right)$ being a leave-one-out estimator, so that it is orthogonal with the error e_i.

This means that $\hat{\beta}$ is asymptotically equivalent to the minimizer of $S_n^*\ \left(\beta\right)$, a NLLS problem. As we know from the Econ710, the asymptotic distribution of the NLLS estimator is identical to least-squares on

$$X_i^* = \frac{\partial}{\partial\beta}G\left(X_i'\beta\right).$$

This implies

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \rightarrow_d N\left(0, V\right)$$

$$\begin{aligned} V &= Q^{-1}\Omega Q^{-1} \\ Q &= \mathrm{E}\left(X_i^* X_i^{*\prime}\right) \\ \Omega &= \mathrm{E}\left(X_i^* X_i^{*\prime} e_i^2\right) \end{aligned}$$

To complete the derivation, we now find this X_i^*.

As $\hat{\beta}$ is $n^{-1/2}$ consistent, we can use a Taylor expansion of $g\left(X_i'\beta_0\right)$ to find

$$g\left(X_i'\beta_0\right) \simeq g\left(X_i'\beta\right) + g^{(1)}\left(X_i'\beta\right)X_i'\left(\beta_0 - \beta\right)$$

where

$$g^{(1)}\left(z\right) = \frac{d}{dz}g\left(z\right).$$

Then

$$\begin{aligned} G\left(X_i'\beta\right) &= \mathrm{E}\left(g\left(X_i'\beta_0\right) \mid X_i'\beta\right) \\ &\simeq \mathrm{E}\Big(g\left(X_i'\beta\right) + g^{(1)}\left(X_i'\beta\right) X_i'\left(\beta_0 - \beta\right) \mid X_i'\beta\Big) \\ &= g\left(X_i'\beta\right) - g^{(1)}\left(X_i'\beta\right) \mathrm{E}\left(X_i \mid X_i'\beta\right)'\left(\beta - \beta_0\right) \end{aligned}$$

since $g\left(X_i'\beta\right)$ and $g^{(1)}\left(X_i'\beta\right)$ are measureable with respect to $X_i'\beta$. Another Taylor expansion for$g\left(X_i'\beta\right)$ yields that this is approximately

$$\begin{aligned} G\left(X_i'\beta\right) &\simeq g\left(X_i'\beta_0\right) + g^{(1)}\left(X_i'\beta\right)\left(X_i - \mathrm{E}\left(X_i \mid X_i'\beta\right)\right)'\left(\beta - \beta_0\right) \\ &\simeq g\left(X_i'\beta_0\right) + g^{(1)}\left(X_i'\beta_0\right)\left(X_i - \mathrm{E}\left(X_i \mid X_i'\beta_0\right)\right)'\left(\beta - \beta_0\right) \end{aligned}$$

the final approximation for β in a $n^{-1/2}$ neighborhood of β_0. (The error is of smaller stochastic order.)

We see that

$$X_i^* = \frac{\partial}{\partial\beta} G\left(X_i'\beta\right) \simeq g^{(1)}\left(X_i'\beta_0\right)\left(X_i - \mathrm{E}\left(X_i \mid X_i'\beta_0\right)\right).$$

Ichimura rigorously establishes this result.

This asymptotic distribution is slightly different than that which would be obtained if the function g were known a priori. In this case, the asymptotic design depends on X_i, not $\mathrm{E}\left(X_i \mid X_i'\beta_0\right)$.

$$Q = \mathrm{E}\left(g^{(1)}\left(X_i'\beta_0\right)^2 X_i X_i'\right)$$

This is the cost of the semiparametric estimation.

Recall when we described identification that we required the dimension of X_i to be 2 or larger. Suppose that X_i is one-dimensional. Then $X_i - \mathrm{E}\left(X_i \mid X_i'\beta_0\right) = 0$ so $Q = 0$ and the above theory is vacuous (as it should be).

The Ichimura estimator achieves the semiparametric efficiency bound for estimation of β when the error is conditionally homoskedastic. Ichimura also considers a weighted least-squares estimator setting the weight to be the inverse of an estimate of the conditional variance function (as in Robinson's FGLS estimator). This weighted LS estimator is then semiparametrically efficient.

8.4 Klein and Spady's Binary Choice Estimator

Klein and Spady (Econometrica, 1993) proposed an estimator of the semiparametric single index binary choice model which has strong similarities with Ichimura's estimator.

The model is

$$y_i = 1\left(X_i'\beta \geq e_i\right)$$

where e_i is an error.

If e_i is independent of X_i and has distribution function g, then the data satisfy the single-index regression

$$\mathrm{E}\left(y \mid x\right) = g\left(x'\beta\right).$$

It follows that Ichimura's estimator can be directly applied to this model.

Klein and Spady suggest a semiparametric likelihood approach. Given g, the log-likelihood is

$$L_n(\beta, g) = \sum_{i=1}^{n} \left(y_i \ln g\left(X_i'\beta\right) + (1 - y_i) \ln\left(1 - g\left(X_i'\beta\right)\right)\right).$$

This is analogous to the sum–of-squared errors function $S_n(\beta, g)$ for the semiparametric regression model.

Similarly with Ichimura, Klein and Spady suggest replacing g with the leave-one-out NW estimator

$$\hat{g}_{-i}\left(X_i'\beta\right) = \frac{\sum_{j \neq i} k\left(\dfrac{\left(X_j - X_i\right)'\beta}{h}\right) y_j}{\sum_{j \neq i} k\left(\dfrac{\left(X_j - X_i\right)'\beta}{h}\right)}.$$

Making this substitution, and adding trimming function, this leads to the feasible likelihood criterion

$$L_n(\beta) = \sum_{i=1}^{n} \left(y_i \ln \hat{g}_{-i}\left(X_i'\beta\right) + (1 - y_i) \ln\left(1 - \hat{g}_{-i}\left(X_i'\beta\right)\right)\right) 1_i(b).$$

Klein and Spady emphasize that the trimming indicator should not be a function of β, but instead of a preliminary estimator. They suggest

$$1_i(b) = 1\left(\hat{f}_{X'\tilde{\beta}}\left(X_i'\tilde{\beta}\right) \geq b\right)$$

where $\tilde{\beta}$ is a preliminary estimator of β, and $\hat{f}$ is an estimate of the density of $X_i'\tilde{\beta}$. Klein and Spady observe that trimming does not seem to matter in their simulations.

The Klein-Spady estimator for β is the value $\hat{\beta}$ which maximizes $L_n(\beta)$.

In many respects the Ichimura and Klein-Spady estimators are quite similar.

Unlike Ichimura, Klein-Spady impose the assumption that the kernel k must be fourth-order (e.g. bias reducing). They also impose that the bandwidth h satisfy the rate $n^{-1/6} < h < n^{-1/8}$, which is smaller than the optimal $n^{-1/9}$ rate for a $4th$ order kernel. It is unclear to me if these are merely technical sufficient conditions, or if there a substantive difference with the semiparametric regression case.

Klein and Spady also have no discussion about how to select the bandwidth. Following the ideas of Hardle, Hall and Ichimura, it seems sensible that it could be selected jointly with β by minimization of $L_n(\beta)$, but this is just a conjecture.

They establish the asymptotic distribution for their estimator. Similarly as in Ichimura, letting

g denote the distribution of e_i, define the function

$$G\left(X_i'\beta\right) = \mathrm{E}\left(g\left(X_i'\beta_0\right) \mid X_i'\beta\right).$$

Then

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \rightarrow_d N\left(0, H^{-1}\right)$$

$$H = \mathrm{E}\left(\frac{\partial}{\partial\beta} G\left(X_i'\beta\right) \frac{\partial}{\partial\beta} G\left(X_i'\beta\right)' \frac{1}{g\left(X_i'\beta_0\right)\left(1 - g\left(X_i'\beta_0\right)\right)}\right)$$

They are not specific about the derivative component, but if I understand it correctly it is the same as in Ichimura, so

$$\frac{\partial}{\partial\beta} G\left(X_i'\beta\right) \simeq g^{(1)}\left(X_i'\beta_0\right)\left(X_i - \mathrm{E}\left(X_i \mid X_i'\beta_0\right)\right).$$

The Klein-Spady estimator achieves the semiparametric efficiency bound for the single-index binary choice model.

Thus in the context of binary choice, it is preferable to use Klein-Spady over Ichimura. Ichimura's LS estimator is inefficient (as the regression model is heteroskedastic), and it is much easier and cleaner to use the Klein-Spady estimator rather than a two-step weighted LS estimator.

8.5 Average Derivative Estimator

Let the conditional mean be

$$\mathrm{E}\left(y \mid x\right) = \mu\left(x\right)$$

Then the derivative is

$$\mu^{(1)}(x) = \frac{\partial}{\partial x}\mu(x)$$

and a weighted average is

$$\mathrm{E}\left(\mu^{(1)}(X)w(X)\right)$$

where $w(x)$ is a weight function. It is particularly convenient to set $w(x) = f(x)$, the marginal density of X. Thus Powell, Stock and Stoker (Econometrica, 1989) define this as the average derivative

$$\delta = \mathrm{E}\left(\mu^{(1)}(X)f(X)\right).$$

This is a measure of the average effect of X on y. It is a simple vector, and therefore easier to report than a full nonparametric estimator.

There is a connection with the single index model, where

$$\mu\left(x\right) = g\left(x'\beta\right)$$

for then

$$\mu^{(1)}(x) = \beta g^{(1)}(x'\beta)$$

$$\delta = c\beta$$

where

$$c = \mathrm{E}\left(g^{(1)}(x'\beta)f(X)\right).$$

Since β is identified only up to scale, the constant c doesn't matter. That is, a (normalized) estimate of δ is an estimate of normalized β.

PSS observe that by integration by parts

$$\begin{aligned}
\delta &= \mathrm{E}\left(\mu^{(1)}(X)f(X)\right) \\
&= \int \mu^{(1)}(x)f(x)^2 dx \\
&= -2\int \mu(x)f(x)f^{(1)}(x)dx \\
&= -2\mathrm{E}\left(\mu(X)f^{(1)}(X)\right) \\
&= -2\mathrm{E}\left(yf^{(1)}(X)\right)
\end{aligned}$$

By the reasoning in CV, an estimator of this is

$$\hat{\delta} = -\frac{2}{n-1}\sum_{i=1}^{n} y_i \hat{f}^{(1)}_{(-i)}(X_i)$$

where $\hat{f}_{(-i)}(X_i)$ is the leave-one-out density estimator, and $\hat{f}^{(1)}_{(-i)}(X_i)$ is its first derivative.

This is a convenient estimator. There is no denominator messing with uniform convergence. There is only a density estimator, no conditional mean needed.

PSS show that $\hat{\delta}$ is $n^{-1/2}$ consistent and asy. normal, with a convenient covariance matrix.

The asymptotic bias is a bit complicated.

Let $q = \dim(X)$. Set $p = ((q+4)/2$ if q is even and $p = (q+3)/2)$ if q is odd. e.g. $p = 2$ for $q = 1$, $p = 3$ for $q = 2$ or $q = 3$ and $p = 4$ for $q = 4$.

PSS require that the kernel for estimation of f be of order at least p. Thus a second-order kernel for $q = 1$, a fourth order for $q = 2$, 3, or 4.

PSS then show that the asymptotic bias is

$$n^{1/2}\left(\mathrm{E}\hat{\delta} - \delta\right) = O\left(n^{1/2}h^p\right)$$

which is $o(1)$ if the bandwidth is selected so that $nh^{2p} \to 0$. This is violated (too big) if h is selected to be optimal for estimation of $\hat{f}$ or $\hat{f}^{(1)}$. This requirement needs the bandwidth to undersmooth to reduce the bias. This type of result is commonly seen in semiparametric methods. Unfortunately, it does not lead to a practical rule for bandwidth selection.

9 Selectivity Models

9.1 Semiparametric Selection Models

The Type-2 Tobit model is the four equation system

$$\begin{aligned} y_{1i}^{*} &= X_{1i}'\beta_1 + u_{1i} \\ y_{2i}^{*} &= X_{2i}'\beta_2 + u_{2i} \\ y_{1i} &= 1\left(y_{1i}^{*} > 0\right) \\ y_{2i} &= y_{2i}^{*} 1\left(y_{1i} = 1\right) \end{aligned}$$

The variables (y_{1i}^{*}, y_{2i}^{*}) are latent (unobserved). The observed variables are $(y_{1i}, y_{2i}, X_{1i}, X_{2i})$. Effectively, y_{2i}^{*} is observed only when $y_{1i} = 1$, equivalently when $y_{1i}^{*} > 0$.

The Type-3 Tobit model is the four equation system

$$\begin{aligned} y_{1i}^{*} &= X_{1i}'\beta_1 + u_{1i} \\ y_{2i}^{*} &= X_{2i}'\beta_2 + u_{2i} \\ y_{1i} &= \max\{y_{1i}^{*}, 0\} \\ y_{2i} &= y_{2i}^{*} 1\left(y_{1i} > 0\right)) \end{aligned}$$

The difference is that y_{1i} is censored rather than binary. We observe y_{2i}^{*} only when there is no censoring.

Typically the second equation is of interest, e.g. the coefficient β_2.

The type-2 model is the classic selection model introduced by Heckman.

It is conventional to assume that the errors (u_{1i}, u_{2i}) are independent of $X_i = (X_{1i}, X_{2i})$.

As you recall from 710, Heckman showed that if we try to estimate β_2 by a regression using the available data, this is estimating the regression of y_{2i} on X_{2i}, conditional on $y_{1i} > 0$, which is

$$\begin{aligned} E\left(y_{2i} \mid X_i, y_{1i} = 1\right) &= X_{2i}'\beta_2 + \mathrm{E}\left(u_{2i} \mid X_i, y_{1i} > 0\right) \\ &= X_{2i}'\beta_2 + \mathrm{E}\left(u_{2i} \mid u_{1i} > -X_{1i}'\beta_1\right) \\ &= X_{2i}'\beta_2 + g\left(X_{1i}'\beta_1\right) \end{aligned}$$

for some function $g(z)$. When (u_{1i}, u_{2i}) are bivariate normal then $g(u)$ is a scaled inverse Mill's ratio. But when the errors are non-normal, the functional form of $g(z)$ is unknown and nonparametric. The one constraint it satisfies is

$$\lim_{z \to \infty} g(z) = \lim_{z \to \infty} \mathrm{E}\left(u_{2i} \mid u_{1i} > -z\right) = \mathrm{E}\left(u_{2i}\right) = 0$$

by normalization.

We can then write the regression for y_{2i} as

$$\begin{aligned} y_{2i} &= X_{2i}'\beta_2 + g\left(X_{1i}'\beta_1\right) + e_i \\ \mathrm{E}\left(e_i \mid X_i, y_{1i} > 0\right) &= 0 \end{aligned}$$

This is a partially linear single index model.

9.2 Two-Step Estimator

This method is developed in a working paper by Powell (1987) and in Li and Wooldridge (Econometric Theory, 2002)

Define $Z_i = X_{1i}'\beta_1$. If Z_i were observed the regression is

$$y_{2i} = X_{2i}'\beta_2 + g\left(Z_i\right) + e_i$$

which is a partially linear model, and can be estimated using Robinson's approach.

For the parially linear model, the intercept is absorbed by g, so it must be excluded from X_{2i}.

Since Z_i is not observed, we can use a two-step approach.

In step 1, β_1 is estimated by a semiparametric estimator, say $\hat{\beta}_1$. (A semiparametric binary choice estimator from the previous section, or a semiparametric Tobit estimator from the next section.) Set $\hat{Z}_i = X_{1i}'\hat{\beta}_1$.

In the second step, β_2 and g are estimated by Robinson's estimator (using the observations for which $y_{1i} = 1$)

Since the second step uses on the generated regressor $\hat{Z}_i = X_{1i}'\hat{\beta}_1$, the asymptotic distribution is affected.

From the text

$$\sqrt{n}\left(\hat{\beta}_2 - \beta_2\right) \rightarrow_d N\left(0, Q^{-1}\left(\Omega_1 + \Psi\Omega_2\Psi'\right)Q^{-1}\right)$$

the covariance terms defined in the text, as typical for two-step estimators.

9.3 Ichimura and Lee's Estimator

Ichimura and Lee (1991) proposed a joint estimator for $\beta = (\beta_1, \beta_2)$ based on the nonlinear regression

$$y_{2i} = X_{2i}'\beta_2 + g\left(X_{1i}'\beta_1\right) + e_i$$

for observations i such that y_{2i} is observed. Thus the first equation is ignored.

Their criterion is

$$S_n\left(\beta\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y_{2i} - X_{2i}'\beta_2 - \hat{g}\left(X_{1i}'\beta_1, \beta_2\right)\right)^2 1_i(b)$$

where $1_i(b)$ is a trimming function

$$\hat{g}\left(X_{1i}'\beta_1,\beta\right)=\frac{\sum_{j\neq i}k\left(\frac{\left(X_{1i}-X_{1j}\right)'\beta_1}{h}\right)\left(y_{2i}-X_{2i}'\beta_2\right)}{\sum_{j\neq i}k\left(\frac{\left(X_{1i}-X_{1j}\right)'\beta_1}{h}\right)}$$

is a leave-one-out NW estimator of $\mathrm{E}\left(y_{2i}-X_{2i}'\beta_2\mid X_{1i}'\beta_1\right)$. (Againl, this is computed only for the observations for which y_{2i} is observed.)

This works, but it ignores the first equation of the system. It is a semiparametric extension of a NLLS Heckit estimator based on the equation

$$y_{2i}=X_{2i}'\beta_2+\sigma_{12}\lambda\left(X_{1i}'\beta_1\right)+e_i.$$

Such estimators ignore the first equation. This is convenient as it simplifies estimation, but ignoring relevant information reduces efficiency. My view is that identification of β_2 versus $g(X_1'\beta_1)$ is based on (dubious) exclusion restrictions plus assuming linearity of the first part.

9.4 Powell's Estimator

An alternative creative estimator was propsoed by Powell (1987, unpublished working paper), reviewed in his chapter 41 of the Handbook of Econometrics.

As for Ichimura and Lee, we ignore the first equation and only consider observations for which y_{2i} are observed.

Take two observations i and j

$$\begin{aligned}y_{2i}&=X_{2i}'\beta_2+g\left(Z_i\right)+e_i\\y_{2j}&=X_{2j}'\beta_2+g\left(Z_j\right)+e_j\end{aligned}$$

and their pairwise difference

$$y_{2i}-y_{2j}=\left(X_{2i}-X_{2j}\right)'\beta_2+g\left(Z_i\right)-g\left(Z_j\right)+e_i-e_j$$

Now focus on observations for which $Z_i\simeq Z_j$. For these observations, $g\left(Z_i\right)-g\left(Z_j\right)\simeq 0$ and β_2 can be estimated by a regression of $y_{2i}-y_{2j}$ on $X_{2i}-X_{2j}$.

This is made operational by using a kernel for Z_i-Z_j, and by replacing Z_i with $\hat{Z}_i=X_{1i}'\hat{\beta}_1$, where $\hat{\beta}_1$ is a first stage estimate of β_1, yielding

$$\begin{aligned}\hat{\beta}_2 &= \left(\sum_i \sum_{j\neq i} k\left(\frac{(X_{1i}-X_{1j})'\hat{\beta}_1}{h}\right)(X_{2i}-X_{2j})(X_{2i}-X_{2j})'\right)^{-1} \\ &\quad \cdot \sum_i \sum_{j\neq i} k\left(\frac{(X_{1i}-X_{1j})'\hat{\beta}_1}{h}\right)(X_{2i}-X_{2j})(y_{2i}-y_{2j})\end{aligned}$$

Unfortunately Powell didn't publish the original paper. This type of estimator effectly identifies β_2 from a small subset of observations, so is unlikely to be precise. The good side is that the nonparametric function g does not need to be estimated in any sense.

9.5 Estimation of the Intercept

While the conditional equation

$$y_{2i} = X_{2i}'\beta_2 + g(Z_i) + e_i$$

excludes an intercept (it is absorbed in g) the original equation of interest

$$y_{2i}^* = \mu + X_{2i}'\beta_2 + u_{2i}$$

say, contains an intercept. Its value can be relevant in practice. That is, the parameters of interest for policy evaluation may be (μ, β_2), not (β_2, g).

To estimate μ, Heckman (AER, 1990) suggest using the observation that the function g satisfies $g(-\infty) = 0$. Thus

$$\mu = \mathrm{E}\left(y_{2i} - X_{2i}'\beta_2 \mid y_{1i} = 1, X_{1i}'\beta_1 = \infty\right) \simeq \mathrm{E}\left(y_{2i} - X_{2i}'\beta_2 \mid y_{1i} = 1, X_{1i}'\beta_1 > \gamma_n\right)$$

where $\gamma_n \to \infty$ is a bandwidth. This can be estimated by

$$\hat{\mu} = \frac{\sum_{i=1}^n 1\left(X_{1i}'\hat{\beta}_1 > \gamma_n\right)(y_{2i} - X_{2i}'\beta_2)}{\sum_{i=1}^n 1\left(X_{1i}'\hat{\beta}_1 > \gamma_n\right)}$$

where the sample is only for those observations for which y_{2i} is observed (those for which $y_{1i} = 1$).

Notet that this estiamtor depends on the first-step estimate $\hat{\beta}_1$.

Andrews and Schafgans (1998, Review of Economic Studies) suggest that a better estimator is obtained by relacing the indicator variable by a DF kernel. They find that the asymptotic distribution has a non-standard rate, depending on the distribution of $X_{1i}'\beta_1$.

10 Censored Models

10.1 Censoring

Suppose that y_i^* is a latent variable, and the observed variable is censored

$$\begin{aligned} y_i &= y_i^* 1\left(y_i^* > 0\right)) \\ &= \begin{cases} 0 & y_i^* \leq 0 \\ y_i^* & y_i^* > 0 \end{cases} \end{aligned}$$

Notationally we have set the censoring point at zero, but this is not essential.

If the distribution of y_i^* is nonparametric, moments of y_i^* are unidentified. We don't observe the full range of support for y_i^*, so anything can happen in that part of the distribution. It follows that moment restrictions are inherently unidentified. When there is censoring, we should be very cautious about moment restriction models.

In contrast, (some) quantiles are identified. Let $q_a(y_i^*)$ and $q_a(y_i)$ denote the α'th quantiles of y_i^* and y_i.

If $P(y_i^* \leq 0) < \alpha$, then $q_\alpha(y_i^*) = q_\alpha(y_i)$. That is, so long as there is less than α percent censored, censoring does not affect quantiles.

Furthermore, if If $P(y_i^* \leq 0) \geq \alpha$ then $q_\alpha(y_i) = 0$. (e.g., if there is 30% censoring, then then quantiles below 30% are identically zero).

This means that we have the relationship:

$$q_\alpha(y_i) = \max(q_\alpha(y_i^* \ ), 0)$$

It follows that we can consistently (and efficiently) estimate quantiles above the α'th on the observed data y_i. These observations lead to the strong conclusion that in the presence of censoring, we should identify parameters through quantile restrictions, not moment restrictions.

Of particular interest is the median $Med\left(y_i^*\right)$. We have

$$Med\left(y_i\right) = \max(Med\left(y_i^*\right), 0)$$

10.2 Powell's CLAD Estimator

The Tobit or censored regression model is

$$\begin{aligned} y_i^* &= X_i'\beta + e_i \\ y_i &= y_i^* 1\left(y_i^* > 0\right)) \end{aligned}$$

The classic Tobit estimator for β is MLE when e_i is independent of X_i and $N(0, \sigma^2)$.

Powell (1984, Journal of Econometrics) made the brilliant observation that when e_i is nonparametric, β is not identified through moment restrictions.

Instead, identify $X_i'\beta$ as the conditional median of y_i, so

$$Med\left(y_i^* \mid X_i\right) = X_i'\beta$$

As we showed in the previous section

$$\begin{aligned} Med\left(y_i \mid X_i\right) &= \max(Med\left(y_i^* \mid X_i\right), 0) \\ &= \max(X_i'\beta, 0) \end{aligned}$$

Thus the conditional median is a specific nonlinear function of the single index $\beta' X_i$.

This shows that the censored observation obeys the nonlinear median regression model

$$\begin{aligned} y_i &= \max(X_i'\beta, 0) + \varepsilon_i \\ Med\left(\varepsilon_i \mid X_i\right) &= 0 \end{aligned}$$

We know that the appropriate method to estimate conditional medians is by least absolute deviations (LAD). This applies as well to nonlinear models. Hence Powell suggested the criterion

$$S_n(\beta) = \sum_{i=1}^{n} \left| y_i - \max(X_i'\beta, 0) \right|$$

or equivalently we can use the criterion

$$S_n(\beta) = \sum_{i=1}^{n} 1\left(X_i'\beta > 0\right) \left| y_i - X_i'\beta \right|.$$

The estimator $\hat{\beta}$ which minimizes $S_n(\beta)$ is called the censored least absolute deviations (CLAD) estimator. The estimator satisfies the asymptotic FOC

$$\sum_{i=1}^{n} 1\left(X_i'\hat{\beta} > 0\right) x_i \operatorname{sgn}\left(y_i - X_i'\hat{\beta}\right) = 0$$

This is the same as the FOC for LAD, but only for the observations for which $X_i'\hat{\beta} > 0$.

Minimization of $S_n(\beta)$ is somewhat more tricky than standard LAD. Bushinsky (PhD dissertation) worked out numerical methods to solve this problem

Powell showed that it has the asymptotic distribution

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \to_d N(0, V)$$

$$\begin{aligned} V &= Q^{-1}\Omega Q^{-1} \\ \Omega &= \mathrm{E}\left(1_i X_i X_i'\right) \\ Q &= 2\mathrm{E}\left(f\left(0 \mid X_i\right) 1_i X_i X_i'\right) \\ 1_i &= 1\left(X_i'\beta > 0\right) \end{aligned}$$

where $f(0 \mid x)$ is the conditional density of e_i given $X_i = x$ at the origin.

The derivation of this result is not much different from that for standard LAD regression. Since the criterion function is not smooth with respect to β, you need to use an empirical process approach, as outlined for example in section 7 of Newey and McFadden's Handbook chapter.

Identification requires that Ω and Q are full rank. This requires that there is not "too much" censoring. As the censoring rate increases, the information in Ω diminishes and precision falls.

10.3 Variance Estimation

When e_i is independent of X_i, then $f(0 \mid x) = f(0)$ and $V = \left(4f\left(0\right)^2\Omega\right)^{-1}$. Practical standard error estimation seems to focus on estimating V under this assumption.

$$\begin{aligned} \hat{V} &= \left(4\hat{f}\left(0\right)^2\hat{\Omega}\right)^{-1} \\ \hat{\Omega} &= \frac{1}{n}\sum_{i=1}^{n} 1\left(X_i'\hat{\beta} > 0\right) X_i X_i' \end{aligned}$$

The difficult part is $f(0)$, in part because $\hat{e}_i$ is only observed for observations with $y_i > 0$.

Hall and Horowitz (1990, Econometric Theory) recommend

$$\hat{f}(0) = \frac{\sum_{i=1}^{n} k\left(\frac{\hat{e}_i}{h}\right) 1\left(y_i > 0\right)}{h\sum_{i=1}^{n} G\left(\frac{X_i'\hat{\beta}}{h}\right)}$$

where $\hat{e}_i = y_i - X_i'\hat{\beta}$, h is a bandwidth, $k(u)$ is a symmetric kernel, and $G(u)$ is integrated kernel. They find that the optimal rate is $h \sim n^{-1/5}$ if k is a second-order kernel, and $h \sim n^{-1/(2\nu+1)}$ if k is a ν'th order kernel. Their paper has an expression for the optimal bandwidth, and discuss possible methods to estimate the bandwidth, but do not present a fully automatic bandwidth method.

An obvious alternative to asymptotic methods is the bootstrap. It is quite common to use bootstrap percentile methods to compute standards errors, confidence intervals, and p-values for LAD, quantile estimation, and CLAD estimation.

10.4 Khan and Powell's Two-Step Estimator

$$S_n(\beta)=\sum_{i=1}^{n}1\left(X_i'\beta>0\right)\left|y_i-X_i'\beta\right|.$$

In this criterion, the coefficient β plays two roles. Khan and Powell (2001, Journal of Econometrics) suggest this double role induces bias in finite samples, and this can be avoided by a two-step estimator.

They suggest first estimating $\tilde{\beta}$ using a semiparametric binary choice estimator, and using this to define the observations for trimming. The second-stage criterion is then

$$S_n(\beta)=\sum_{i=1}^{n}1\left(X_i'\tilde{\beta}>0\right)\left|y_i-X_i'\beta\right|.$$

(In the theoretical treatment, the indictor function is replaced with a smooth weighting function, but they claim this is only to make the theory easy, and they use the indicator function in their simulations.) The second-stage estimator minimizes this criterion, which is just LAD on the trimmed sub-sample. Khan and Powell argue that this two-step estimator falls in the class of Andrews' MINPIN estimators, so the asymptotic distribution is identical to Powell's estimator.

10.5 Newey and Powell's Weighted CLAD Estimator

When e_i is not independent of X_i, the asymptotic covariance matrix of the CLAD estimator suggests that it is inefficient and can be improved. Newey and Powell (1990, Econometric Theory) compute the semiparametric efficiency bound, and find that it is attained by the estimator minimizing the weighted criterion

$$\begin{aligned}S_n(\beta)&=\sum_{i=1}^{n}w_i\left|y_i-\max(X_i'\beta,0)\right|\\ w_i&=2f\left(0\mid X_i\right)\end{aligned}$$

The estimator $\hat{\beta}$ which minimizes this criterion is a weighted CLAD estimator, and the authors show that it has the asymptotic distribution

$$\sqrt{n}\left(\hat{\beta}-\beta\right)\to_d N(0,V)$$

$$V=\left(4\mathrm{E}\left(f\left(0\mid X_i\right)^2 1_i X_i X_i'\right)\right)^{-1}$$

The conditional density plays a similar role to the conditional variance for GLS regression.

This efficiency result is general to median regression, not just censored regression. That is, the weighted LAD estimator achieves the asymptotic efficiency bound for median regression. The unweighted estimator is efficient when $f(0\mid x)=f(0)$ (essentially, when e_i is independent of X_i).

Feasible versions of this estimator are challenging to construct. Newey and Powell suggest a method based on nearerst neighbor regression estimation of the conditional distribution function.

I don't know if this has been noticed elsewhere, but here is a useful observation.

Suppose that the error e_i only depends on x_i through a scale effect. That is

$$e_i = \sigma(X_i)z_i$$

where z_i is independent of X_i, with density $f_z(z)$ and median zero. Then the conditional density of e_i given $X_i = x$ is

$$f(e \mid x) = \frac{1}{\sigma(x)} f_z\left(\frac{e}{\sigma(x)}\right)$$

so at the origin

$$f(0 \mid x) = \frac{1}{\sigma(x)} f_z(0)$$

Thus the optimal weighting is $w_i \sim \sigma(X_i)^{-1}$, which takes the same form as in the case of GLS in regression. The interpretation of $\sigma(x)$ is a bit different (it is identified on median restrictions).

10.6 Nonparametric Censored Regression

The models discussed in the previous sections assume that the conditional median is linear in X_i – a highly parametric assumption. It would be desireable to extend the censored regression model to allow for nonparametric median functions. A nonparametric model would take the form

$$\begin{aligned} Med\left(y_i^* \mid X_i\right) &= g\left(X_i\right) \\ y_i &= y_i^* 1\left(y_i^* > 0\right)) \end{aligned}$$

with g nonparametric. The conditional median for the observed dependent variable is

$$Med\left(y_i \mid X_i\right) = \max(g\left(X_i\right), 0).$$

We can define the conditional median function

$$g^*(x) = \max(g\left(x\right), 0).$$

Since $g(x)$ is nonparametric then so is $g^*(x)$, although it does satisfy $g^*(x) \geq 0$.

A feasible approach to estimate $g^*(x)$ is to simply use standard nonparametric median regression. It is unclear if any information is lost in ignoring the censoring. My only thought is that function $g^*(x)$ will typically have a "kink" at $g(x) = 0$, and this is smoothed over by nonparametric methods, which suggests inefficiency.

An alternative suggestion is Lewbel and Linton (Econometrica, 2002). They impose the strong assumption that the error $e_i = y_i^* - g(X_i)$ is independent of X_i, and develop nonparametric estimates of $g(x)$ using kernel methods.

11 Nearest Neighbor Methods

11.1 kth Nearest Neighbor

An alternative nonparametric method is called k-nearest neighbors or k-nn. It is simiar to kernel methods with a random and variable bandwidth. The idea is to base estimation on a fixed number of observations k which are closest to the desired point.

Suppose $X \in \mathbb{R}^q$ and we have a sample $\{X_1, ..., X_n\}$.

For any fixed point $x \in \mathbb{R}^q$, we can calculate how close each observation X_i is to x using the Euclidean distance $\|x\| = (x'x)^{1/2}$. This distance is

$$D_i = \|x - X_i\| = \left((x - X_i)'(x - X_i)\right)^{1/2}$$

This is just a simple calculation on the data set.

The order statistics for the distances D_i are $0 \leq D_{(1)} \leq D_{(2)} \leq \cdots \leq D_{(n)}$.

The observations corresponding to these order statistics are the "nearest neighbors" of x. The first nearest neighbor is the observation closest to x, the second nearest neighbor is the observation second closest, etc.

This ranks the data by how close they are to x. Imagine drawing a small ball about x and slowly inflating it. As the ball hits the first observation X_i, this is the "first nearest neighbor" of x. As the ball further inflates and hits a second observation, this observation is the second nearest neighbor.

The observations ranked by the distances, or "nearest neighbors", are $\{X_{(1)}, X_{(2)}, X_{(3)}, ..., X_{(n)}\}$.

The k'th nearest neighbor of x is $X_{(k)}$.

For a given k, let

$$R_x = \left\|X_{(k)} - x\right\| = D_{(k)}$$

denote the Euclidean distance between x and $X_{(k)}$. R_x is just the k'th order statistic on the distances D_i.

Side Comment: When X is multivariate the nearest neighbor ordering is not invariant to data scaling. Before applying nearest neighbor methods, is therefore essential that the elements of X be scaled so that they are similar and comparable across elements.

11.2 k-nn Density Estimate

Suppose $X \in \mathbb{R}^q$ has multivariate density $f(x)$ and we are estimating $f(x)$ at x.

A multivariate uniform kernel is

$$w(\|u\|) = c_q^{-1} 1 \left(\|u\| \leq 1\right)$$

where

$$c_q = \frac{\pi^{q/2}}{\Gamma\left(\dfrac{q+2}{2}\right)}$$

is the volume of unit ball in $\mathbb{R}^q$. If $q = 1$ then $c_1 = 2$.

Treating R_x as a bandwidth and using this uniform kernel

$$\begin{aligned}\tilde{f}(x) &= \frac{1}{nR_x^q}\sum_{i=1}^n c_q^{-1}1\left(\|x - X_i\| \leq R_x\right) \\ &= \frac{1}{nR_x^q}\sum_{i=1}^n c_q^{-1}1\left(D_i \leq R_x\right)\end{aligned}$$

But as $R_x = D_{(k)}$ is the k'th order statistic for D_i, there are precisely k observations where $\|x - X_i\| \leq R_x$. Thus the above equals

$$\tilde{f}(x) = \frac{k}{nR_x^q c_q}$$

To compute $\tilde{f}(x)$, all you need to know is R_x.

The estimator is inversely proportional to R_x. Intuitively, if R_x is small this means that there are many observations near x, so $f(x)$ must be large, while if R_x is large this means that there are not many observations near x, so $f(x)$ must be small.

A motivation for this estimator is that the effective number of observations to estimate $\tilde{f}(x)$ is k, which is constant regardless of x. This is in contrast to the conventional kernel estimator, where the effective number of observations varies with x.

While the traditional k-nn estimator used a uniform kernel, smooth kernels can also be used. A smooth k-nn estimator is

$$\tilde{f}(x) = \frac{1}{nR_x^q}\sum_{i=1}^n w\left(\frac{\|x - X_i\|}{R_x}\right)$$

where w is a kernel weight function such that

$$\int_{\mathbb{R}^q} w\left(\|u\|\right)(du) = 1.$$

In this case the estimator does not simplify to a function of R_x only

The analysis of k-nn estimates are complicated by the fact that R_x is random.

The solution is to calculate the bias and variance of $\hat{f}(x)$ conditional on R_x, which is similar to treating R_x as fixed. It turns out that the conditional bias and variance are identical to those of the standard kernel estimator:

$$\begin{aligned}E\left(\tilde{f}(x) \mid R_q\right) &\simeq f(x) + \frac{\kappa_2(w)\nabla^2 f(x)R_x^2}{2} \\ var\left(\tilde{f}(x) \mid R_q\right) &\simeq \frac{R\left(w\right)f(x)}{nR_x^q}.\end{aligned}$$

We can then approximate the unconditional bias and variance by taking expectations:

$$\begin{aligned} E\left(\tilde{f}(x)\right) &\simeq f(x)+\frac{\kappa_2(w)\nabla^2 f(x)}{2}E\left(R_x^2\right) \\ var\left(\tilde{f}(x)\right) &\simeq \frac{R\left(w\right)f(x)}{n}E\left(R_x^{-q}\right) \end{aligned}$$

We see that to evaluate these expressions we need the moments of $R_x = D_{(k)}$ the k'th order statistic for D_i. The distribution function for order statistics is well known. Asymptotic moments for the order statistics were found by Mack and Rosenblatt (Journal of Multivariate Analysis, 1979):

$$E\left(R_x^{\lambda}\right) \simeq \left(\frac{k/n}{c_q f\left(x\right)}\right)^{\lambda/q}$$

This depends on the ratio k/n and the density $f(x)$ at x. Thus

$$\begin{aligned} E\left(R_x^2\right) &\simeq \left(\frac{k}{nc_q f\left(x\right)}\right)^{2/q} \\ E\left(R_x^{-q}\right) &\simeq \frac{c_q f\left(x\right)n}{k} \end{aligned}$$

Substituting,

$$\begin{aligned} Bias\left(\tilde{f}(x)\right) &\simeq \frac{\kappa_2(w)\nabla^2 f(x)}{2}\left(\frac{k}{nc_q f\left(x\right)}\right)^{2/q} \\ &= \frac{\kappa_2(w)\nabla^2 f(x)}{2\left(c_q f\left(x\right)\right)^{2/q}}\left(\frac{k}{n}\right)^{2/q} \end{aligned}$$

$$\begin{aligned} var\left(\tilde{f}(x)\right) &\simeq \frac{R\left(w\right)f(x)}{n}\frac{c_q f\left(x\right)}{k}n \\ &= \frac{R\left(w\right)c_q f\left(x\right)^2}{k} \end{aligned}$$

For k-nn estimation, the integer k is similar to the bandwidth h for kernel density estimation, except that we need $k \to \infty$ and $k/n \to 0$ as $n \to \infty$.

The MSE is of order

$$MSE\left(\tilde{f}(x)\right) = O\left(\left(\frac{k}{n}\right)^{4/q}+\frac{1}{k}\right)$$

This is minimized by setting

$$k \sim n^{4/(4+q)}.$$

The optimal rate for the MSE is

$$MSE\left(\tilde{f}(x)\right) = O\left(n^{-4/(4+q)}\right)$$

which is the same as for kernel density estimation with a second-order kernel.

Kernel estimates $\hat{f}$ and k-nn estimates $\tilde{f}$ behave differently in the tails of $f(x)$ (where $f(x)$ is small). The contrast is

$$\begin{aligned} Bias\left(\hat{f}(x)\right) &\simeq \nabla^2 f(x) \\ Bias\left(\tilde{f}(x)\right) &\simeq \frac{\nabla^2 f(x)}{f\left(x\right)^{2/q}} \end{aligned}$$

$$\begin{aligned} var\left(\hat{f}(x)\right) &\simeq f\left(x\right) \\ var\left(\tilde{f}(x)\right) &\simeq f\left(x\right)^2 \end{aligned}$$

In the tails, where $f(x)$ is small, $\tilde{f}(x)$ will have larger bias but smaller variance than $\hat{f}(x)$. This is because the k-nn estimate uses more effective observations than the kernel estimator. It is difficult to rank one estimator versus the other based on this comparison. Another way of viewing this is that in the tails $\tilde{f}(x)$ will tend to be smoother than $\hat{f}(x)$.

11.3 Regression

Nearest neighbor methods are more typically used for regression than for density estimation.

The regression model is

$$\begin{aligned} y_i &= g\left(X_i\right) + e_i \\ E\left(e_i \mid X_i\right) &= 0 \end{aligned}$$

The classic k-nn estimate of $g(x)$ is

$$\tilde{g}(x) = \frac{1}{k}\sum_{i=1}^{n} 1\left(\|x - X_i\| \leq R_x\right) y_i$$

This is the average value of y_i among the observations which are the k nearest neighbors of x.

A smooth k-nn estimator is

$$\tilde{g}(x) = \frac{\sum_{i=1}^n w\left(\dfrac{\|x - X_i\|}{R_x}\right) y_i}{\sum_{i=1}^n w\left(\dfrac{\|x - X_i\|}{R_x}\right)},$$

a weighted average of the k nearest neighbors.

The asymptotic analysis is the same as for density estimation. Conditional on R_x, the bias and variance are approximately as for NW regression. The conditional bias is proportional to R_x^2 and the variance to $1/nR_x^q$. Taking unconditional expecations and using the formula for the moments of R_x give expressions for the bias and variance of $\tilde{g}(x)$. The optimal rate is $k \sim n^{4/(4+q)}$ and the optimal convergence rate is the same as for NW estimation.

As for density estimation, in the tails of the density of X, the bias of the k-nn estimator is larger, and the variance smaller, than the NW estimator $\hat{g}(x)$. Since the effective number of observations k is held constant across x, $\tilde{g}(x)$ is smoother than $\hat{g}(x)$ in the tails.

11.4 Local Linear k-nn Regression

As pointed out by Li and Racine, local linear esitmation can be combined with the nearest neighbor method.

A simple estimator (corresonding to a uniform kernel) is to take the k observations "nearest" to x, and fit a linear regression of y_i on X_i using these observations.

A smooth local linear k-nn estimator fits a weighted linear regression

11.5 Cross-Validation

To use nearest neighbor methods, the integer k must be selected. This is similar to bandwidth selection, although here k is discrete, not continuous.

K.C. Li (Annals of Statistics, 1987) showed that for the $k-$nn regression estimator under conditional homoskedasticity, it is asymptotically optimal to pick k by Mallows, Generalized CV, or CV. Andrews (Journal of Econometrics, 1991) generalized this result to the case of heteroskedasticity, and showed that CV is asymptotically optimal. The CV criterion is

$$CV(k) = \sum_{i=1}^{n} \left(y_i - \tilde{g}_{-i}(X_i)\right)^2$$

and $\tilde{g}_{-i}(X_i)$ is the leave-one-out k-nn estimator of $g(X_i)$. The method is to select k by minimizing $CV(k)$. As k is discrete, this amounts to computing $CV(k)$ for a set of values for k, and finding the minimizing value.

12 Series Methods

12.1 General Approach

A model has parameters (β, η) where β is finite-dimensional and η is nonparametric. (Sometimes, there is no β.) We will focus on regression.

The function η is approximated by a series – a finite dimensional model which depends on an integer K and a K dimensional parameter θ. Let $\eta_K(\theta)$ denote this approximating function.

Typically, the parameters (β, θ) are estimated by a conventional parametric technique $(\hat{\beta}, \hat{\theta})$. Then $\hat{\eta} = \eta_L(\hat{\theta})$

Tasks:

- To find a class of functions $\eta_K(\theta)$ which are good approximations to η.
- Study the bias (due to the finite dimensional approximation) and variance of the estimators
- Find optimal rates for K to diverge to infinity
- Find rules for selection of K
- Show that $\hat{\beta}, \hat{\eta}$ are asymptotically normal.
- Asymptotic variance computation, and standard error calculation.

Data Transformation: Typically the methods are applied after transforming the regressors X to lie in a specific compact space, such as $[0, 1]$.

12.2 Regression and Splines

Take the univariate regression

$$y_i = g(X_i) + e_i$$

In this case, $\eta = g$.

Series Approximations:

- power series (polynomial)
  - works for low order polynomials
  - unstable for high order polynomials
- trigonometric (sin and cos functions)
  - bounded functions
  - can produce "wiggly" implausible nonparametric function estimates
- splines

- piecewise polynomial of order r
- continuous derivatives up to $r-1$
- cubic splines popular
- join points (knots) can be selected evenly, or estimated

12.3 Splines

It is useful to define the "positive part" function

$$
\begin{aligned}
(a)_+ &= \max[0, a] \\
&= \begin{cases} 0 & a < 0 \\ a & a \geq 0 \end{cases}
\end{aligned}
$$

Linear, quadratic and cubic splines with knots at $t_1 < t_2 < \cdots < t_{J-1}$ are

$$g_K(x) = \theta_0 + \theta_1 x + \sum_{j=1}^{J-1} \theta_{1+j} (x - t_j)_+$$

$$g_K(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \sum_{j=1}^{J-1} \theta_{2+j} (x - t_j)_+^2$$

$$g_K(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \sum_{j=1}^{J-1} \theta_{3+j} (x - t_j)_+^3$$

This model is set up so that it is everywhere a polynomial of order s, with continuous derivatives of order up to s, and the s''th derivative changing discontinuously at the knots. Cubic splines are smooth approximating functions, flexible, and popular. The approximation improves as the number of knots increases. The dimension of θ is $K = J + s$.

For a given set of knots the function g_K is linear in the parameters. Define

$$z = z(x) = \begin{pmatrix} 1 & x & x^2 & x^3 & (x - t_1)_+^3 & \cdots & (x - t_{J-1})_+^3 \end{pmatrix}',$$

then

$$g_K(x) = \theta_K' z$$

12.4 B Splines

Another popular class of series approximation are called B-splines. These are basis functions which are bounded, integrable and density-shaped. They can be constructed from a variety of basic shapes. Polynomials are common.

Let $X \in [0,1]$ and divide the support into J equal subintervals, with knots are $t_j = j/J$, $j = 0, 1, ..., J$. We also need knots outside of $[0,1]$ so let $t_j = j/m$ for all integers j.

An r'th order B-spline is a piecewise $(r-1)$-order polynomial.

A linear $(r = 2)$ B-spline base functions are linear on two adjacent subintervals, zero elsewhere. They take the form

$$B_2\left(x \mid t_j, t_{j+1}, t_{j+2}\right) = \left(x - t_j\right)_+ - 2\left(x - t_{j+1}\right)_+ + \left(x - t_{j+2}\right)_+ .$$

A quadratic $(r = 3)$ B-spline base function is piecewise quadratic over three subintervals

$$B_3\left(x \mid t_j, t_{j+1}, t_{j+2}, t_{j+3}\right) = \left(x - t_j\right)_+ - 3\left(x - t_{j+1}\right)_+ + 3\left(x - t_{j+2}\right)_+ - \left(x - t_{j+3}\right)_+ .$$

For general r

$$B_r\left(x \mid t_j, \cdots, t_{j+r}\right) = \sum_{s=0}^{r} (-1)^s \binom{r}{s} \left(x - t_{j+s}\right)_+ .$$

The B-spline is a linear combination of these basis functions.

$$\begin{aligned} g_K(x) &= \sum_{j=1-r}^{J-1} \theta_j B_r\left(x \mid t_j, \cdots, t_{j+r}\right) \\ &= \theta_K' z \end{aligned}$$

where $z = z(x)$ is the vector of the basic functions. The dimension of θ is $K = J + r + 1$

12.5 Estimation

For all of the examples, the function g_K is linear in the parameters (at least if the knots are fixed). Define the vector $Z_i = z(X_i)$ as the sample base function transformations. For example, in the case of a cubic spline

$$Z_i = \left(\begin{array}{cccccccc} 1 & X_i & X_i^2 & X_i^3 & (X_i - t_1)_+^3 & \cdots & (X_i - t_{J-1})_+^3 \end{array} \right)' .$$

From Z_i, construct the regressor matrix Z. The LS estimate of θ_K is $\hat{\theta}_K = Z\left(Z'Z\right)^{-1} Z'y$. The estimate of $g(x)$ is $\hat{g}(x) = z'\hat{\theta}_K$, that of $g(X_i)$ is $\hat{g}(X_i) = z_i'\hat{\theta}_K$ and that of the vector $g = (g(X_1), ..., g(X_n))'$ is

$$\hat{g} = Z\hat{\theta}_K = Py$$

where

$$P = Z\left(Z'Z\right)^{-1} Z'$$

is a projection matrix.

12.6 Bias

Since $y = g + e$ then

$$\begin{aligned} E\left(\hat{\theta}_K \mid X\right) &= \left(Z'Z\right)^{-1} Z'E\left(y \mid X\right) \\ &= \left(Z'Z\right)^{-1} Z'g \\ &= \theta_K^* \end{aligned}$$

the coefficient from a regression of g on Z. This is the effective projection or pseudo-true value.

Similarly,

$$E\left(\hat{g} \mid X\right) = Pg = g_K^*$$

is the projection of g on Z.

The bias in estimation of g is

$$E\left(\hat{g} - g \mid X\right) = g_K^* - g.$$

If the series approximation works well, the bias will decrease as K gets increases. If g is α-times differentiable, then for splines and power series

$$\sup_x \left|g_K^*(x) - g(x)\right| \leq O\left(K^{-\alpha}\right).$$

The integrated squared bias is

$$ISB_K = \int \left(g_K^*(x) - g(x)\right)^2 dF(x) \leq O\left(K^{-2\alpha}\right)$$

where $F(x)$ is the marginal distribution of X.

This is approximately the same as the empirical average

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\left(g_K^*(X_i) - g(X_i)\right)^2 &= \frac{1}{n}\left(g_K^* - g\right)'\left(g_K^* - g\right) \\ &= \frac{1}{n} g'\left(I - P'\right)\left(I - P\right) g \\ &= \frac{1}{n} g'\left(I - P\right) g \end{aligned}$$

12.7 Integrated Squared Error

The integrated squared error of $\hat{g}(x)$ for $g(x)$ is

$$\begin{aligned} ISE &= \int (\hat{g}(x) - g(x))^2 \, dF(x) \\ &\simeq \frac{1}{n} \sum_{i=1}^{n} (\hat{g}(X_i) - g(X_i))^2 \\ &= \frac{1}{n} (\hat{g} - g)' (\hat{g} - g) \end{aligned}$$

Since

$$\begin{aligned} \hat{g} - g &= P(g + e) - g \\ &= Pe - (I - P) g \end{aligned}$$

then

$$\begin{aligned} ISE_K &= \frac{1}{n} (Pe - (I - P) g)' (Pe - (I - P) g) \\ &= \frac{1}{n} e'PPe + \frac{1}{n} g' \left(I - P'\right) (I - P) g - \frac{2}{n} e'P (I - P) g \end{aligned}$$

and when P is a projection matrix (as for LS estimation) then this simplifies to

$$ISE_K = \frac{1}{n} e'Pe + ISB_K \qquad (1)$$

The first part represents estimation variance, the second is the integrated squared bias.

If the error is conditionally homoskedastic, then the conditional expectation of the first part is

$$\begin{aligned} E\left(\frac{1}{n} e'Pe \mid X\right) &= \frac{1}{n} \operatorname{tr}\left(PE\left(ee' \mid X\right)\right) \\ &= \frac{1}{n} \operatorname{tr}(P) \sigma^2 \\ &= \frac{K}{n} \sigma^2 \end{aligned}$$

In general, it can be shown that

$$\frac{1}{n} e'Pe = O_p\left(\frac{K}{n}\right)$$

Put together with the analyis of the ISB, we have

$$ISE_K \le O_p\left(\frac{K}{n}\right) + O\left(K^{-2\alpha}\right).$$

The optimal rate for K is $K = n^{1/(2\alpha+1)}$ yielding a MSE convergence $n^{-2\alpha/(2\alpha+1)}$. This is the same as the best rate attained by kernel regression using higher-order kernels or local polynomials.

12.8 Asymptotic Normality

The dimension of $\hat{\theta}_K$ grows with n, so we do not discuss its asymptotic distribution.

At any x, the estimate of $g(x)$ is $\hat{g}(x) = z'\hat{\theta}_K$, a linear function of the OLS estimator $\hat{\theta}_K$. Let $\hat{V}_K$ be the conventional (White) asymptoticcovariance matrix estimator for $\hat{\theta}_K$, so that for $z'\hat{\theta}_K$ is $z'\hat{V}_K z$. Applying the CLT we can find

$$\frac{\sqrt{n}\left(\hat{g}(x) - g_K^*(x)\right)}{\sqrt{z'\hat{V}_K z}} \rightarrow_d N(0,1)$$

Since the estimator is nonparametric, it is biased, so the estimator should be centered at the projection or pseudo-true value rather than the true $g(x)$. Alternative, if K is larger than optimal, so the estimator is "undersmoothed", then the squared bias will be of smaller order than the variance and it can be omitted from the asymptotic expression.

The bottom line is that for series estimation, we calculate standard errors using the conventional formula, as if the model were parametric. However, it is not constructive to focus on standard errors for individual coefficients, as they do not have individual meaning. Rather, standard errors should be for identifiable parameters, such as the conditional mean $g(x)$.

12.9 Selection of Series Terms

The role of K is similar to that of the bandwidth in kernel regression. Automatic data-dependent procedures are necessary for implementation.

As we worked out before, the integrated squared error is

$$ISE_K = \frac{1}{n} e'Pe + ISB_K$$

The optimal K minimizes this expression, but it is unknown.

We can estimate it using the sum-of-squared residuals from a model. For a given K, there regressors define a projection matrix P, fitted value $\hat{g} = Py$ and residual vector $\hat{e}_K = y - Py$. Note that

$$\begin{aligned} \hat{e}_K &= (I-P)\,y \\ &= (I-P)\,g + (I-P)\,e \end{aligned}$$

Thus the SSE is

$$\begin{aligned}
\frac{1}{n}\hat{e}_K'\hat{e}_K &= \frac{1}{n}g'(I-P)g + \frac{2}{n}g'(I-P)e + \frac{1}{n}e'(I-P)e \\
&= ISB_K - \frac{1}{n}e'Pe + \frac{2}{n}g'(I-P)e + \frac{1}{n}e'e \\
&= ISE_K - \frac{2}{n}e'Pe + \frac{1}{n}2g'(I-P)e + \frac{1}{n}e'e
\end{aligned}$$

Taking expectations conditional on X,

$$\begin{aligned}
E\left(\frac{1}{n}\hat{e}_K'\hat{e}_K \mid X\right) &= E(ISE_K \mid X) - E\left(\frac{2}{n}e'Pe \mid X\right) + \sigma^2 \\
&= E(ISE_K \mid X) - \frac{2K\sigma^2}{n} + \sigma^2
\end{aligned}$$

where the second line holds under conditional homoskedasticity.

Thus $\frac{1}{n}\hat{e}'\hat{e}$ is biased for ISE_K, but this can be corrected if we correct for the bias. This leads to Mallows (1973) criteria

$$C_K = \hat{e}_K'\hat{e}_K + 2K\hat{\sigma}^2$$

where $\hat{\sigma}^2$ is a preliminary estimate of σ^2. The scale doesn't matter, so I have multiplied through by n as is conventional, and the final σ^2 term doesn't matter, as it is independent of K.

The Mallows estimate $\hat{K}$ is the value which minimizes C_K.

A method which does not require homoskedasticity is cross-validation. The CV criterion is

$$CV_K = \sum_{i=1}^{n}\left(y_i - \hat{g}_{-i}^{K}(X_i)\right)^2$$

where $\hat{g}_{-i}^{K}$ is a K-th order series estimator omitting observation i. The CV estimate $\hat{K}$ is the value which minimizes CV_K.

Li (1987, Annals of Statistics) showed under quite minimal conditions that Mallows, GCV, and CV are asymptotically optimal for selection of K, in the sense that

$$\frac{ISE_{\hat{K}}}{\inf_k ISE_K} \to_p 1$$

Andrews (1991, JoE) showed that this optimality only extends to the heteroskedastic case if CV is used for selection. The reason is that the Mallows criterion uses homoskedasticity to calculate the bias adjustment, as we showed above, and this is not needed under CV.

12.10 Partially Linear and Additive Models

Suppose

$$y_i = W_i'\gamma + g(X_i) + e_i$$

with g nonparametric. A series approximation for g is $z'\theta_K$ yielding the model for estimation

$$y_i = W_i'\gamma + z_i'\theta_K + error_i$$

which is estimated by least-squares. The estimate for γ is similar to that from the Robinson kernel estimator, which had a residual-regression interpretation.

The asymptotic distribution for $\hat{\gamma}$ is the same as for the Robinson estimator, under the condition that the nonparametric component has MSE converging faster than $n^{-1/2}$, e.g. if $K/n + K^{-2\alpha} = o\left(n^{-1/2}\right)$. This is similar to the requirement for the Robinson estimator.

You can easily generalize this idea to multiple additive nonparametric components

$$y_i = W_i'\gamma + g_1\left(X_{1i}\right) + g_2\left(X_{2i}\right) + e_i$$

In practice, the components X_{1i} and X_{2i} are real-valued.

As discussed in Li-Racine, W_i can contain nonlinear interaction effects between X_{1i} and X_{2i}, such as $X_{1i}X_{2i}$. The main requirement is that the components of W_i cannot be additively separable in X_{1i} and X_{2i}. So in this sense the additive model can allow for simple interaction effects.

13 Endogeneity and Nonparametric IV

13.1 Nonparametric Endogeneity

A nonparametric IV equation is

$$\begin{aligned} Y_i &= g(X_i) + e_i \qquad (1) \\ E(e_i \mid Z_i) &= 0 \end{aligned}$$

In this model, some elements of X_i are potentially endogenous, and Z_i is exogenous.

We have studied this model in 710 when g is linear. The extension to nonlinear g is not obvious.

The first – and primary – issue is identification. What does g mean?

Let

$$\lambda(z) = E(Y_i \mid Z_i = z)$$

and take the conditional expectation of (1):

$$\begin{aligned} \lambda(z) &= E(Y_i \mid Z_i = z) \\ &= E(g(X_i) + e_i \mid Z_i = z) \\ &= E(g(X_i) \mid Z_i = z) \\ &= \int g(x) f(x \mid z)\, dx. \end{aligned}$$

The functions $\lambda(z)$ and $f(x \mid z)$ are identified. The unknown nonparametric $g(x)$ is a solution to the integral equation

$$\lambda(z) = \int g(x) f(x \mid z)\, dx.$$

The difficulty is that the solution $g(x)$ is not necessarily unique. The mathematical problem is that the solution g is not necessarily continuous in the function f. The non-uniqueness of g is called the "ill-posed inverse problem".

A solution is to restrict the space of allowable functions g. For example, the linear model $g(x) = x'\beta$ is linear, then the above equation reduces to

$$\begin{aligned} \lambda(z) &= \beta' \int x f(x \mid z)\, dx \\ &= \beta' E(X_i \mid Z_i = z). \end{aligned}$$

Idenfication of β in the linear model exploits this simple relationship.

13.2 Newey-Powell-Vella's Triangular Simultaneous Equations

Newey, Powell, Vella (1989, Econometrica)

The model is

$$\begin{aligned} Y_i &= g(X_i) + e_i \qquad (2) \\ E(e_i \mid Z_i) &= 0 \end{aligned}$$

plus a reduced-form equation for X_i

$$\begin{aligned} X_i &= \Pi(Z_i) + u_i \qquad (3) \\ E(u_i \mid Z_i) &= 0 \end{aligned}$$

Thus $\Pi(z)$ is the conditional mean of X_i given $Z_i = z$. The vectors X_i and Z_i may overlap, so X_i can contain both endogenous and exogenous variables.

NPV then take expectations of (2) given X_i and Z_i

$$\begin{aligned} E(Y_i \mid X_i, Z_i) &= E(g(X_i) + e_i \mid X_i, Z_i) \\ &= g(X_i) + E(e_i \mid X_i, Z_i) \end{aligned}$$

Since X_i is endogenous, the latter conditional expectation is not zero. In general, this cannot be simplified further. But NPV observe the following. From (3), X_i is a function of Z_i and u_i, so conditioning on X_i and Z_i is equivalent to conditioning on u_i and Z_i. Hence

$$E(Y_i \mid X_i, Z_i) = g(X_i) + E(e_i \mid u_i, Z_i)$$

Next, suppose that Z_i is strongly exogenous in the sense that

$$E(e_i \mid u_i, Z_i) = E(e_i \mid u_i) = g_2(u_i)$$

That is, conditional on u_i, Z_i provides no information about the mean of the error e_i. In this case we have the simplification

$$E(Y_i \mid X_i, Z_i) = g(X_i) + g_2(u_i)$$

which implies

$$\begin{aligned} Y_i &= g(X_i) + g_2(u_i) + \varepsilon_i \\ E(\varepsilon_i \mid u_i, X_i) &= 0 \end{aligned}$$

This is an additive regression model, with the regressor u_i unobserved but identified.

Is g identified?

Since (2) is a (reduced-form) regression, Π is identified. Thus u_i is identified.

Then the functions g and g_2 are identified so long as X_i and u_i are distinct. NPV discussed several identification conditions. One is:

Theorem: If there is no functional relationship between x and u, then $g(x)$ is identified up to an additive constant.

The additive constant qualitification is required in all additive nonparametric models.

The authors propose the following series estimator:

1. Estimate $\hat{\Pi}_L(z) = \hat{\theta}'_L Z_{Li}$ using a series in Z_i with L terms, say
   (a) $\hat{\theta}_L = (Z'_L Z_L)^{-1} Z'_L X$ where Z_L are the basic functions of Z
   (b) Residual $\hat{u}_i = X_i - \hat{\Pi}(X_i)$
2. Create a basis transformation for X_i, and a separate one for $\hat{u}_i$, with K coefficients.
   (a) spline functions of X_i
   (b) spline functions of $\hat{u}_i$,
3. Least-squares regression of Y_i on these basis functions, obtain $\hat{g}$ and $\hat{g}_2$

NPV show that this estimator is consistent and asymptotically normal. The conditions require that the functions g and Π be sufficiently smooth (enough derivatives), and that the number of terms K and L diverge to infinity in a controlled way. The regularity conditions are not particularly helpful.

It is not clear how K and L should be selected in practice. A reasonable suggestion is to select L by cross-validation on the reduced form regression, and then select K by cross-validation on the second-stage regression. The trouble is that the two stages are not orthogonal, so the MSE of the second stage is affected by the first stage, so it is unlikely that the CV criterion will correctly reflect this.

We are primarily interested in the estimate $\hat{g}$ of g (the structural form equation). The estimates of $(\hat{g}, \hat{g}_2)$ in the second stage are asymptotically normal, but affected by the first stage (the generated regressors problem).

One solution is to write out the correct asymptotic covariance matrix for the two-step estimator as discussed in NPV

The other, easier approach is to view the problem as a one-step estimator. Stack the moment equations from each step. Then the two-step estimates are equivalent to a one-step estimator – just-identified GMM on the stacked equations. The covariance matrix may be calculated for the estimates using the standard GMM formula.

The authors include an application to wage/hours profile.

13.3 Newey and Powell's Estimator

Newey and Powell (Econometrica, 2003) propose a nonparametric method which avoids the strong exogeneity assumption, but imposes restrictions on allowable g.

Return to the base model

$$\begin{aligned} Y_i &= g(X_i) + e_i \\ E(e_i \mid Z_i) &= 0 \end{aligned}$$

and tthe integral equation

$$E(Y_i \mid Z_i) = \int g(x) f(x \mid Z_i) \, dx$$

To identify g, NP point out that one solution is to assume that g lives in a compact space. Their paper is based on this assumption, and impose this on their estimates of g.

Next, suppose that $g(x)$ can be approximated using a series approximation. Write this as

$$g(x) \simeq g_K(x) = \gamma_K' p_K(x)$$

where γ_K is a K vector of parameters and $p_K(x)$ is a K vector of basis functions. Compactness of g can be impose by assuming that γ_K is bounded. They use $\gamma_K' W_K \gamma_K \leq C$ where W_K is a specific weight matrix and C is a pre-determined constant.

Substituting the series expasion into the integral equation,

$$\begin{aligned} E(Y_i \mid Z_i) &\simeq \gamma_K' \int p_K(x) f(x \mid Z_i) \, dx \\ &= \gamma_K' E(p_K(X_i) \mid Z_i) \\ &= \gamma_K' h_K(Z_i) \end{aligned}$$

where

$$h_K(z) = E(p_K(X_i) \mid Z_i = z)$$

is the K vector of conditional expectations of the basis function transformations of X_i.

We thus have the regression models

$$\begin{aligned} Y_i &= \gamma_K' h_K(Z_i) + v_i \\ E(v_i \mid Z_i) &= 0 \end{aligned}$$

and

$$\begin{aligned} p_K(X_i) &= h_K(z) + \eta_i \\ E(\eta_i \mid Z_i) &= 0 \end{aligned}$$

NW suggest a two-step estimator.

1. Select the basis functions $p_K(x)$
2. Non-parametrically regress each element of $p_K(x)$ on Z_i using series methods. The estimates are collected in to the vector $\hat{h}_K(z)$.
3. Regress Y_i on $\hat{h}_K(z)$ (least squares) to obtain $\hat{\gamma}_K$.
4. The estimate of interest is $\hat{g}(x) = \hat{\gamma}_K' p_K(x)$

Identification requires that g (and thus $\hat{g}$) satisfy a compactness condition. NW recommend that this be imposed on $\hat{g}$ by restricting the estimate $\hat{\gamma}_K$ to satisfy $\hat{\gamma}_K' W_K \hat{\gamma}_K \leq C$. (This can be easily imposed by constrained LS regression.) As the constant C is arbitrary it is unclear what this means in practice.

NW discuss conditions for consistency of $\hat{g}$.

The estimator $\hat{\gamma}_K$ is a two-step estimator in the generate regressor class. Thus the conventional standard errors for $\hat{\gamma}_K$ (and thus $\hat{g}$) are incorrect.

The method is extended and applied to Engel curve estimation by Blundell, Chen and Kristensen (Econometrica, 2007). They extend the analysis of identification, and include a proof of asymptotic normality, how to calculate standard errors, and computational implication issues.

Other important related papers are referenced include Hall and Horowitz (Annals of Statistics, 2005), and Darolles, Florens and Renault, “Nonparametric Instrumental Regression” (early version 2002, current version 2009, still unpublished).

This general topic is clearly very important to econometrics and underdeveloped.

13.4 Ai and Chen (Econometrica, 2003)

Take the conditional moment restriction model

$$E\left[\rho\left(Z, \alpha_0\right) \mid X\right] = 0$$

(Notationally, we have switched the Z and X from the previous section, and I do this to follow Li-Racine, who simply followed the notation in the original papers.) Here, Z is the “endogenous” variables, and X are exogenous. The function ρ is a residual (or moment) equation). E.g. $\rho(Z) = y - z_1'\alpha$ in a linear framework). The are interested in the semiparametric framework in which $\alpha = (\theta, g)$ where θ is parametric and g is nonparametric. Their focus is on efficient estimation of the parametric component θ. (In this sense their work takes a different focus from the papers earlier reviewed, which focused on the nonparametric component.)

An example is the partially linear regression model with an endogenous regressor, where the focus is on the parametric component.

For the moment consider estimation of α assuming that it is parametric

Define the conditional mean and variance of $\rho(Z,\alpha)$ for generic values of α :

$$\begin{aligned} m(x,\alpha) &= E\left[\rho(Z,\alpha) \mid X=x\right] \\ \sigma^2(x) &= var\left[\rho(Z,\alpha) \mid X=x\right] \end{aligned}$$

Note that at the true value α_0

$$m(x,\alpha_0)=0$$

for all x.

If the functions m and σ^2 were known and α were parametric, a reasonable estimator for α would be found by minimizing the squared error criterion

$$\sum_{i=1}^{n} \frac{m(X_i,\alpha)^2}{\sigma^2(X_i)}.$$

For simplicity, suppose $\sigma^2(x)=1$, then the criterion simplifies to

$$\sum_{i=1}^{n} m(X_i,\alpha)^2 = m(\alpha)' m(\alpha)$$

where $m(\alpha)$ is the vector of stacked $m(X_i,\alpha)$.

As m is unknown this is infeasible. We can replace m with an estimate.

For any fixed α estimate m by a series regression. That is, approximate

$$m(x,\alpha) \simeq p_K(x)'\pi_K(\alpha)$$

where $p_K(x)$ is a K vector of basis functions in x. Then

$$m(\alpha) = P\pi_K(\alpha)$$

where P is the matrix of the regressors $p_K(X_i)$.

Let $\rho(\alpha)$ be the vector of stacked $\rho(Z_i,\alpha)$.

We estimate $\pi_K(\alpha)$ by LS of $\rho(\alpha)$ on P :

$$\hat{\pi}_K(\alpha) = \left(P'P\right)^{-1} P'\rho(\alpha)$$

The estimate of $m(\alpha)$ is

$$\hat{m}(\alpha) = P\hat{\pi}_K(\alpha) = P\left(P'P\right)^{-1} P'\rho(\alpha)$$

and the squared error criterion is estimated by

$$\begin{aligned}\hat{m}(\alpha)'\hat{m}(\alpha) &= \rho(\alpha)'P\left(P'P\right)^{-1}P'P\left(P'P\right)^{-1}P'\rho(\alpha)\\ &= \rho(\alpha)'P\left(P'P\right)^{-1}P'\rho(\alpha)\end{aligned}$$

which is a GMM criterion. If α were parametric, it could be estimated by minimizing this criterion. Indeed this is conventional GMM using the instrument set P under the assumption of homoskedasticity.

Now, as $\alpha = (\theta, g)$ includes a nonparametric component, Ai and Chen suggest replacing g by a series approximation:

$$g(z) \simeq q_L(z)'\beta_L$$

where $q_L(z)$ is an L vector of basis functions in z.

The moment equation

$$\rho(z, \theta, g(z)) \simeq \rho\left(z, \theta, q_L(z)'\beta_L\right)$$

is then a function of the parameters θ and β_L. For fixed (θ, β_L) define the $n \times 1$ vector $\rho(\theta, \beta_L)$ of stacked elements $\rho\left(Z_i, \theta, q_L(Z_i)'\beta_L\right)$. Replacing $\rho(\alpha)$ with $\rho(\theta, \beta_L)$ we have the revised GMM criterion

$$\begin{aligned}&\rho(\theta,\beta_L)'P\left(P'P\right)^{-1}P'\rho(\theta,\beta_L)\\ = &\sum_{i=1}^n \rho\left(Z_i,\theta,q_L(Z_i)'\beta_L\right)p_K(X_i)\left(\sum_{i=1}^n p_K(X_i)p_K(X_i)'\right)^{-1}\sum_{i=1}^n p_K(X_i)\rho\left(Z_i,\theta,q_L(Z_i)'\beta_L\right)\end{aligned}$$

The estimates $(\hat{\theta}, \hat{\beta}_L)$ then minimize this function.

This is not Ai and Chen's preferred estimator. For efficiency, they suggest estimating $\hat{\sigma}^2(X_i)$, the conditional variance, and using the weighted criterion.

This criterion is

$$\sum_{i=1}^n \frac{\left(p_K(X_i)'\left(P'P\right)^{-1}P'\rho(\theta,\beta_L)\right)^2}{\hat{\sigma}^2(X_i)} = \rho(\theta,\beta_L)'P\left(P'P\right)^{-1}\left(P'\hat{D}^{-1}P\right)\left(P'P\right)^{-1}P'\rho(\theta,\beta_L).$$

This is GMM with the eficient weight matrix $\left(P'P\right)^{-1}\left(P'\hat{D}^{-1}P\right)\left(P'P\right)^{-1}$.

Ai and Chen demonstrate that the estimate $\hat{\theta}$ is root-n asymptotically normal and asymptotically efficient (in the semiparametric sense)

14 Time Series

14.1 Stationarity

A (multivariate) time series y_t is an $m \times 1$ vector observed over time $t = 1, ..., n$. We think of the sample as a "window" out of an infinite past and infinite future.

A series y_t is strictly stationary if the joint distribution $(y_t, ..., y_{t+h})$ is constant across t for all h. An implication of stationary is that any finite moment is time-invariant.

A linear measure of dependence is autocovariance

$$\begin{aligned} \gamma(k) &= cov\left(y_t, y_{t-k}\right) \\ &= E\left(\left(y_t - Ey_t\right)\left(y_{t-k} - Ey_{t-k}\right)'\right) \end{aligned}$$

Stationarity implies that $\gamma(k)$ is constant over time t.

A loose definition of ergodicity is that $\gamma(k) \to 0$ as $k \to \infty$. A rigorous definition requires a measure-theoretic treatment, but intuitively that history is independent of the infinite past.

Simple time series models are built using fundamental "shocks" e_t, typically normalized so that $Ee_t = 0$.

(1) The most basic shock e_t is iid.

(2) Martingale Difference Sequence (MDS). e_t is a MDS if $E\left(e_t \mid e_{t-1}, e_{t-2}, ...\right) = 0$

(3) White noise. If $Ee_te_{t-h} = 0$ for all $h \geq 1$.

An iid shock is a MDS, and a MDS is white noise, but the reverse is not true. An iid shock is unpredictable. A MDS is unpredictable in the mean. A white noise shock is linearly unpredictable.

A process is m-dependent if y_t and y_{t+k} are independent for $k > m$. In this case, $\gamma(k) = 0$ for $k > m$.

A simple model of time dependence is a moving average. A MA(1) is

$$y_t = e_t + \theta e_{t-1}$$

with e_t white noise. You can calculate that $\gamma(1) = \theta\sigma_e^2$. An MA(q) is q-dependent.

Another simple model is an autoregression. An AR(1) is

$$y_t = \alpha y_{t-1} + e_t$$

where e_t is iid. The series is stationary if $|\alpha| < 1$. You can calculate that $\sigma_y^2 = \gamma(1) = \sigma_e^2/(1-\alpha^2)$ and $\gamma(k) = \alpha^k\sigma_y^2$. An AR process is not m-depenent.

An example of a MDS which is not iid is an ARCH process

$$\begin{aligned} e_t &= \sigma_t z_t \\ \sigma_t^2 &= \omega + \alpha e_{t-1}^2 \end{aligned}$$

or a GARCH(1,1)

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha e_{t-1}^2$$

where z_t is an iid shock. This series is predictable in the square (in the variance) but not in the mean.

An example of a white noise process which is not an MDS is the nonlinear MA

$$y_t = e_t + e_{t-1}e_{t-2}$$

You can calculate that this is white noise. Yet $E\left(e_t \mid e_{t-1}, e_{t-2}, ...\right) = e_{t-1}e_{t-2} \neq 0$. This is a nonlinear process.

14.2 Time Series Averages

Let $\theta = Ey_t$ be estimated by

$$\hat{\theta} = \frac{1}{n}\sum_{t=1}^{n} y_t$$

If y_t is stationary,

$$E\hat{\theta} = Ey_t = \theta$$

so the estimator is unbiased.

The variance of the standardized estimator is

$$\begin{aligned} var(\sqrt{n}\left(\hat{\theta}-\theta\right)) &= \frac{1}{n}E\left(\sum_{t=1}^{n}\left(y_t-\theta\right)\right)\left(\sum_{t=1}^{n}\left(y_t-\theta\right)\right)' \\ &= \frac{1}{n}E\left(\sum_{t=1}^{n}\sum_{j=1}^{n}\left(y_t-\theta\right)\left(y_j-\theta\right)'\right) \\ &= \frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{n}\gamma\left(t-j\right) \end{aligned}$$

This can be simplified to equal

$$\sum_{k=-(n-1)}^{n-1}\left(\frac{n-k}{n}\right)\gamma(j)$$

Notice that it is bounded by the inequality

$$var(\sqrt{n}\left(\hat{\theta}-\theta\right)) \le \sum_{k=-\infty}^{\infty}|\gamma(j)|$$

Hence, if $\sum_{k=0}^{\infty}|\gamma(j)| < \infty$, then $var(\sqrt{n}\left(\hat{\theta}-\theta\right))$ is bounded. By Markov's inequality, it follows that $\hat{\theta} \to_p \theta$. This is a rather simple proof of consistency for time-series averages. This proof uses stronger conditions than necessary.

The Ergodic Theorem states that if y_t is strictly stationary and ergodic, and $E\left|y_t\right| < \infty$, then $\hat{\theta} \to \theta$ a.s.

For a central limit theorem, we need a stronger summability condition on the covariances. As we showed above,

$$\begin{aligned} var(\sqrt{n}\left(\hat{\theta} - \theta\right)) &= \sum_{k=-(n-1)}^{n-1} \left(\frac{n-k}{n}\right)\gamma(j) \\ &\to \sum_{k=-\infty}^{\infty} \gamma(j) \\ &\equiv \Omega \end{aligned}$$

as $n \to \infty$. Thus the asymptotic variance of $\hat{\theta}$ is Ω, not $var(y_t)$. Under regularity conditions the estimator is asymptotically normal

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \to_d N\left(0, \Omega\right)$$

The variance Ω is sometimes called the "long-run covariance matrix" and is also a scale of the spectral density of y_t at frequency zero.

The fact that Ω involves an infinite sum of covariances means that standard error calculations for time-series averages needs to take this into account. Estimation of Ω is called "HAC" estimation in econometrics (for heteroskedasticity and autocorrelation consistent covariance matrix estimation), and is often called the "Newey-West estimator" due to an early influential paper (Econometrica, 1987) by Whitney Newey and Ken West.

14.3 GMM

This carries over to GMM estimation. If θ is the solution to

$$Em\left(y_t, \theta\right) = 0$$

where m is a known function, the GMM estimator minimizes a quadratic funtion in the sample moment of $m(y_t, \theta)$ to find $\hat{\theta}$. The asymptotic distribution of $\hat{\theta}$ is determined by the sample average of $m(y_t, \theta_0)$ at the true value θ_0, so if these are autocorrelated, then the asymptotic distribution of $\hat{\theta}$ will involve the long-run covariance matrix.

Theorem 1 *Under general regularity conditions the GMM estimator for stationary time-series data satisfies*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\right)$$

where

$$\boldsymbol{G} = \mathrm{E}\frac{\partial}{\partial \boldsymbol{\theta}'} m(y_t, \theta_0)$$

and

$$\boldsymbol{\Omega} = \sum_{k=-\infty}^{\infty} \mathrm{E}\left(\boldsymbol{m}_t \boldsymbol{m}_{t-k}'\right).$$

A important simplication occurs when $m(y_t, \theta_0)$ is serially uncorrelated. This occurs in dynamically well-specified models or correctly-specified MLE. In this case, $m(y_t, \theta_0)$ will be a MDS, serially uncorrelated, and thus

$$\boldsymbol{\Omega} = \mathrm{E}\left(\boldsymbol{m}_t \boldsymbol{m}_t'\right).$$

14.4 Linear Regression

Consider linear regression

$$y_t = \theta' x_{t-1} + e_t$$

This includes autoregressions, as x_{t-1} can include $y_{t-1}, y_{t-2}, ...$

In this model, the LS estimator of θ is GMM, using the moment condition

$$m(y_t, x_{t-1}, \theta) = x_{t-1}\left(y_t - \theta' x_{t-1}\right)$$

Note that at the true θ_0,

$$m(y_t, x_{t-1}, \theta_0) = x_{t-1} e_t$$

The properties of m_t depend on the properties of e_t.

If the model is a true regression, then $\theta' x_{t-1}$ is the conditional mean, and the error is conditionally mean zero and thus a MDS:

$$E\left(e_t \mid I_{t-1}\right) = 0$$

where I_{t-1} contains all lagged information. In this case m_t is a MDS as well:

$$E\left(m_t \mid I_{t-1}\right) = E\left(x_{t-1} e_t \mid I_{t-1}\right) = x_{t-1} E\left(e_t \mid I_{t-1}\right) = 0$$

Thus the LS estimator is asymptotically normal, with a conventional covariance matrix.

On the other hand, if the model is an approximation, a linear projection, then $\theta' x_{t-1}$ is not necessarily the conditional mean, so e_t is not necessarily a MDS. Then m_t will not necessarily be serially uncorrelated, and Ω will contain the autocovariances of m_t

14.5 Density Estimation

Suppose y_t is univariate and strictly stationary with marginal distribution $F(y)$ with density $f(y)$. The kernel density estimator of $f(y)$ is

$$\hat{f}(y) = \frac{1}{nh} \sum_{t=1}^{n} k\left(\frac{y - y_t}{h}\right)$$

where $k(u)$ is a kernel function and h is a bandwidth.

As the function is linear, the expectation is not affected by time-series dependence. That is,

$$E\hat{f}(y) = E\frac{1}{h}k\left(\frac{y-y_t}{h}\right)$$

and this is the same as in the cross-section case. Hence the bias of $\hat{f}(y)$ is unchanged by dependence.

To calculate the variance, for simplicity assume that y_t is m-depenent. The variance is

$$\begin{aligned}
var(\hat{f}(y) &= E\left(\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{h}k\left(\frac{y-y_t}{h}\right) - E\frac{1}{h}k\left(\frac{y-y_t}{h}\right)\right)\right)^2 \\
&= \frac{1}{n^2}\sum_{t=1}^{n}\sum_{j=1}^{n} cov\left(\frac{1}{h}k\left(\frac{y-y_t}{h}\right), \frac{1}{h}k\left(\frac{y-y_j}{h}\right)\right) \\
&= \frac{1}{n}\sum_{k=-(n-1)}^{n-1}\left(\frac{n-k}{n}\right)\frac{1}{h^2}cov\left(k\left(\frac{y-y_t}{h}\right), k\left(\frac{y-y_{t-k}}{h}\right)\right) \\
&= \frac{1}{n}\sum_{k=-m}^{m}\left(\frac{n-k}{n}\right)\frac{1}{h^2}cov\left(k\left(\frac{y-y_t}{h}\right), k\left(\frac{y-y_{t-k}}{h}\right)\right) \\
&\simeq \frac{1}{nh^2}Ek\left(\frac{y-y_t}{h}\right)^2 + \frac{2}{n}\sum_{k=1}^{m}\frac{1}{h^2}Ek\left(\frac{y-y_t}{h}\right)k\left(\frac{y-y_{t-k}}{h}\right)
\end{aligned}$$

The second-to-last step uses m-dependence. The first part in the final line is the same as in the cross section case, and is asymptotically $\dfrac{R(k)f(y)}{nh}$.

Now take the second part in the final line, which is the sum of the m components. Let $f(u_0, u_k)$ be the joint density of (y_t, y_{t-k}) for $k > 0$. Then

$$\begin{aligned}
\frac{1}{h^2}Ek\left(\frac{y-y_t}{h}\right)k\left(\frac{y-y_{t-k}}{h}\right) &= \int\int\frac{1}{h^2}k\left(\frac{y-u_0}{h}\right)k\left(\frac{y-u_k}{h}\right)f\left(u_0,u_k\right)du_0du_k \\
&= \int\int k\left(v_0\right)k\left(v_k\right)f\left(y-hv_0, y-hv_k\right)dv_0dv_k
\end{aligned}$$

where I made two change-of-variables, $u_0 = y - hv_0$ and $u_k = y - hv_k$, which has Jacobian h^2. Expanding the joint density and integrating, this equals

$$\int k\left(v_0\right)dv_0\int k\left(v_k\right)dv_k f\left(y,y\right) + o(1) = f\left(y,y\right)$$

the joint density, evaluated at (y, y). We have found that

$$\begin{aligned}
var(\hat{f}(y) &= \frac{R(k)f(y)}{nh} + \frac{2m}{n}f\left(y,y\right) + o\left(\frac{1}{n}\right) \\
&\simeq \frac{R(k)f(y)}{nh}
\end{aligned}$$

which is dominated by the same term as in the cross-section case. Time-series dependence has no effect on the asymptotic bias and variance of the kernel estimator! This is in strong contrast to the parametric case, where correlation affects the asymptotic variance.

The technical requirement is that the joint density of the observations (y_t, y_{t-k}) is smooth at (y, y). In the cross-section case we only required smoothness of the marginal density. In the time series case we need smoothness of the joint densities, even though we are just estimating a marginal density.

The intuition is that kernel (nonparametric) estimation is averaging the data locally in the $y-$dimension, where there is no time-series dependence, not in the time-dimension. That is, $\hat{f}(y)$ is a average of observations y_t that are close to y. This subset of observations are not necessarily close to each other in time. Thus they have low joint dependence, and the contribution of joint dependence to the asymptotic variance is of small order relative to the nonparametric smoothing.

The theoretical implication of this result is that the theory of nonparametric kernel density estimation carries over from the iid case to the time-series case without essential modification. (However, proofs of the theorems requires attention to dependence and stronger smoothness conditions.)

14.6 One-Step-Ahead Point Forecasting

Given information up to n, we want a forecast f_{n+1} of y_{n+1}. What does this mean? We want f_{n+1} to be "close" to the realized y_{n+1}. Let $L(f, y)$ denote the loss associated with a forecast f of realized y. The risk of the estimator is the expected loss $R(f \mid I_n) = E\left(L(f, y) \mid I_n\right).$ A convenient loss function is quadratic $L(f, y, I_n) = (f - y)^2$ in which case the best forecast is the conditional mean $f = E\left(y_{n+1} \mid I_n\right).$ In this sense it is common for a point forecast for y_{n+1} to be an estimate of the conditional mean.

Let $x_{t-1} = (y_{t-1}, y_{t-2}, ..., y_{t-q})$, and assume that the conditional mean of y_t is (approximately) a function only of x_{t-1}. Then

$$\begin{aligned} y_t &= g\left(x_{t-1}\right) + e_t \\ E\left(e_t \mid I_{t-1}\right) &= 0 \end{aligned}$$

(One issue in nonparametrics is letting $q \to \infty$ as $n \to \infty$ so to allow for infinite dependence.) Then the optimal point forecast (given quadratic loss) for y_{n+1} is $g(x_n)$. A feasible point forecast is $\hat{g}(x_n)$ where $\hat{g}(x)$ is an estimator of g.

The conventional linear approach is to set $\hat{g}(x) = x'\hat{\theta}$ where $\hat{\theta}$ is LS of y on X. This imposes a linear model.

Any other nonparametric estimator can be used. In particular, a local linear estimator nests the linear (AR) model as a special case, but allows for nonlinearity.

14.7 One-Step-Ahead Interval Forecasting

An interval forecast is $\hat{C} = [f_1, f_2]$. The goal is select $\hat{C}$ so that $P(y_{n+1} \in \hat{C}) = .9$ (or some other pre-specified coverage probability) and so that $\hat{C}$ is as short as possible. Often people don't mention the goal of a short $\hat{C}$. But this is critical, otherwise we can design silly $\hat{C}$ with the desired coverage. For example, let $\hat{C} = \mathbb{R}$ with probability .9, and $\hat{C} = \hat{g}(x_n)$ with probability .1. This has exact coverage .9 but is clearly not using sample information intelligently.

A desired forecast inteval sets f_1 and f_2 equal to the .05 and .95 quantiles of the conditional distribution of y given x.

We discussed this problem earlier. Given an estimate $\hat{g}(x)$ for the conditional, set $\hat{e}_t = y_t - \hat{g}(x_{t-1})$ and estimate the CDF $\hat{F}(e \mid x)$, conditional distribution of $\hat{e}_t$ given x_{t-1}. (As a special case, you can use the unconditional DF $\hat{F}(e)$.) Let $\hat{q}_\alpha(x)$ be the α quantile of this conditional distribution. Then setting $\hat{f}_1 = \hat{g}(x_n) + \hat{q}_{.05}(x_n)$ and $\hat{f}_2 = \hat{g}(x_n) + \hat{q}_{.95}(x_n)$ the forecast interval is $\hat{C} = [\hat{f}_1, \hat{f}_2]$.

An alternative method to display forecast uncertainty is to use a density forecast. Let $f(y \mid x)$ denote the conditional density of y_t given x_{t-1}. An estimate of this conditional density is

$$\hat{f}(y \mid x) = \hat{f}_e\left(y - \hat{g}(x) \mid x\right)$$

where $\hat{f}_e$ is a (kernel) estimate of the conditional density of $\hat{e}_t$ given x_t. (As a special case you can use the unconditional density estimate, which is approximating the error e_t as being independent of x_{t-1}). The forecast density is then

$$\hat{f}(y \mid x_n) = \hat{f}_e\left(y - \hat{g}(x_n) \mid x_n\right)$$

14.8 Multi-Step-Ahead Forecasting

Practical forecasts are typically for multiple periods out of sample. Let $h \geq 1$ be the forecast horizon (a positive integer). It is convenient to switch notation and write the problem as forecasting y_{t+h} given x_t.

Given squared error loss, the optimal forecast is the conditional mean $f = E\left(y_{n+h} \mid I_n\right)$.

There are two main approaches to multi-step forecasting.

One method, the direct approach, is to specify a model for the h-step conditional mean, e.g.

$$\begin{aligned} y_{t+h} &= g\left(x_t\right) + e_{t+h} \\ E\left(e_{t+h} \mid I_t\right) &= 0 \end{aligned}$$

This is estimated by a (parametric or nonparametric) regression of y_{t+h} on x_t. The forecast is then $f = \hat{g}(x_n)$.

This requires a different "model" for each forecast horizon h, and could even imply different selected x_t for different horizons.

The strengths of this approach is that it directly models the object of interest. It is believed to be relatively robust to misspecification. One weakness is that the error e_{t+h} cannot be uncorrelated whe $h > 1$. It is necessarily a MA$(h-1)$ process. This might invalidate conventional order selection methods (this is an open question)

The other method, called the "iterated" or "plug-in" approach, is to estimate a one-step-ahead forecast, and iterate. When $g(x) = x'\theta$ is linear and the goal is point forecasting, this is relatively simple, as iterated mean forecasts are functions only of θ. But for forecast intervals or when $g(x)$ is nonlinear, θ is insufficient. The only method is to estimate the one-step-ahead distribution

$$\hat{F}(y \mid x) = \hat{F}_e\left(y - \hat{g}(x) \mid x\right)$$

(or density), and iterate the entire one-step-ahead distribution. As this involves h-fold integration, it is easiest accomplished through simulation.

If $\hat{F}_e$ is estimated by NW, it can be written as

$$\hat{F}_e\left(e \mid x\right) = \sum_{t=1}^{n} p_t(x) 1\left(\hat{e}_t \leq e\right)$$

where

$$p_t(x) = \frac{K\left(H^{-1}\left(X_t - x\right)\right)}{\sum_{j=1}^{n} K\left(H^{-1}\left(X_j - x\right)\right)}$$

This is a discrete distribution, with probability mass $p_t(x)$ at $\hat{e}_t$. Thus to draw from $\hat{F}_e\left(e \mid x\right)$ is simply to draw from the empirical distribution of the residuals $\hat{e}_t$, with the weighted probabilities $p_t(x)$. (The simpliest case treats the error as independent of x_t, in which case $p_t(x) = n^{-1}$ for all t.)

[If a smoothed distribution estimator is used, then a simulation draw from the kernel is added.]

To make a draw from $\hat{F}(y \mid x)$, draw e^* from $\hat{F}_e(e \mid x)$ as described above, and then set $y^* = \hat{g}(x) + e^*$. Then y^* has the conditional distribution $\hat{F}(y \mid x)$.

To make multi-step-ahead draws:

1. Given x_n, draw e_{n+1} from $\hat{F}_e(e \mid x_n)$ and set $y_{n+1} = \hat{g}(x_n) + e_{n+1}$.

2. Define $x_{n+1} = (y_{n+1}, x_n)$.

3. Given x_{n+1}, draw e_{n+2} from $\hat{F}_e(e \mid x_{n+1})$ and set $y_{n+2} = \hat{g}(x_{n+1}) + e_{n+2}$.

4. Iterate until you attain y_{n+h}.

This creates one draw from the estimated h-step-ahead distribution, say $y_{b,n+h}$. Repeat for $b = 1, ..., B$, similar to bootstrapping.

The point forecast, the conditional mean, is the expected value of the conditional distribution, so is estimated by the average of the simulations

$$f_h = \frac{1}{B} \sum_{b=1}^{B} y_{b,n+h}.$$

A 90% forecast interval is constructed from the 5% and 95% quantiles of the $y_{b,n+h}$. A forecast density can be calculated by applying a kernel density estimator to $y_{b,n+h}$.

This procedure actually creates a joint distribution on the multi-step-ahead conditional distribution $(y_{n+1},...,y_{n+h})$. Perhaps this could be constructively used for some purpose. (e.g., what is the probability that unemployment rate will remain above 7% for all of the next 12 months?)

The acknowledged good feature of the iterative method (in parametric models) is that it is more accurate when the one-step-ahead distribution is correctly specified. The ackowledged downside is that it is is believed to be non-robust to misspecification. Errors in the one-step-ahead distribution can be magnifide when iterated multiple times. It is unclear how these statements apply to the non-parametric setting. The models are both correctly specified (as the bandwidth decreases the models become more accurate) yet are explcitly misspecified (in finite samples, any fitted nonparametric model is incomplete and biased).

14.9 Model Selection

Information criterion are widely used to select linear forecasting models.

Suppose the K'th model has K parameters and residual variance $\hat{\sigma}_K^2 = n^{-1}\sum \hat{e}_t^2$ where $\hat{e}_t = y_t - \hat{\theta}' x_{t-1}$

Popular methods in econometrics include AIC, BIC

$$\begin{aligned} AIC_K &= n\ln\left(\hat{\sigma}_k^2\right) + 2K \\ BIC_K &= n\ln\left(\hat{\sigma}_k^2\right) + \ln(n)K \end{aligned}$$

Less popular in econometrics, but widely used in time-series more generally, is Predictive Least Squares (PLS) introduced by Rissanen

$$\begin{aligned} PLS_K &= \sum_{t=P}^{n} \tilde{e}_t^2 \\ \tilde{e}_t &= y_t - x_{t-1}'\tilde{\theta}_{t-1} \\ \tilde{\theta}_{t-1} &= \left(\sum_{j=1}^{t-1} x_{j-1}x_{j-1}'\right)^{-1} \sum_{j=1}^{t-1} x_{j-1}y_j \end{aligned}$$

This is a time-series genearlization of CV. $\tilde{e}_t$ is a predictive residual. The sequential estimate $\tilde{\theta}_{t-1}$ uses observations up to $t-1$ for a one-step forecast. The PLS criterion is the sum of squared out-of-sample prediction errors. The PLS criterion needs a start-up sub-sample P before evaluating the first residual. Unfortunately the criterion can be sensitive to P, and there is no good guide for its selection.

While PLS is not typically used in econometrics for explicit model selection, it is very commonly used for model comparison. Models are frequently compared by so-called "out of sample performance". In practice, this involves comparing the PLS criterion across models. When this is done, it is frequently described as if this is an “objective” comparison of performance. In fact, it is just a comparison of the PLS criterion. While a good criterion, it is not necessarily superior to other criterion, and is not at all infallible as a model selection criterion.

It is widely asserted that these methods can be applied to the direct multi-step-ahead context. I am skeptical, as the proofs are typically omitted, and the multi-step-ahead model has correlated errors. This may be worth investigating.

15 Model Selection

15.1 KLIC

Suppose a random sample is $\mathbf{y} = y_1, , ..., y_n$ has unknown density $\mathbf{f}(\mathbf{y}) = \prod f(y_i)$.

A model density $\mathbf{g}(\mathbf{y}) = \prod g(y_i)$.

How can we assess the "fit" of $\mathbf{g}$ as an approximation to $\mathbf{f}$?

One useful measure is the Kullback-Leibler information criterion (KLIC)

$$KLIC(\mathbf{f}, \mathbf{g}) = \int \mathbf{f}(\mathbf{y}) \log \left(\frac{\mathbf{f}(\mathbf{y})}{\mathbf{g}(\mathbf{y})} \right) d\mathbf{y}$$

You can decompose the KLIC as

$$\begin{aligned} KLIC(\mathbf{f}, \mathbf{g}) &= \int \mathbf{f}(\mathbf{y}) \log \mathbf{f}(\mathbf{y}) d\mathbf{y} - \int \mathbf{f}(\mathbf{y}) \log \mathbf{g}(\mathbf{y}) d\mathbf{y} \\ &= C_f - E \log \mathbf{g}(\mathbf{y}) \end{aligned}$$

The constant $C_f = \int \mathbf{f}(\mathbf{y}) \log \mathbf{f}(\mathbf{y}) d\mathbf{y}$ is independent of the model g.

Notice that $KLIC(f, g) \geq 0$, and $KLIC(f, g) = 0$ iff $g = f$. Thus a "good" approximating model g is one with a low KLIC.

15.2 Estimation

Let the model density $g(y, \theta)$ depend on a parameter vector θ. The negative log-likelihood function is

$$\mathcal{L}(\theta) = -\sum_{i=1}^{n} \log g(y_i, \theta) = -\log \mathbf{g}(\mathbf{y}, \theta)$$

and the MLE is $\hat{\theta} = \text{argmin}_\theta \mathcal{L}(\theta)$. Sometimes this is called a "quasi-MLE" when $g(y, \theta)$ is acknowledged to be an approximation, rather than the truth.

Let the minimizer of $-E \log g(y, \theta)$ be written θ_0 and called the pseudo-true value. This value also minimizes $KLIC(f, g(\theta))$. As the likelihood divided by n is an estimator of $-E \log g(y, \theta)$, the MLE $\hat{\theta}$ converges in probability to θ_0. That is,

$$\hat{\theta} \rightarrow_p \theta_0 = \underset{\theta}{\text{argmin}}\, KLIC(f, g(\theta))$$

Thus QMLE estimates the best-fitting density, where best is measured in terms of the KLIC.

From conventional asymptotic theory, we know

$$\sqrt{n} \left(\hat{\theta}_{QMLE} - \theta_0 \right) \rightarrow_d N(0, V)$$

$$
\begin{aligned}
V &= Q^{-1}\Omega Q^{-1} \\
Q &= -E\frac{\partial^2}{\partial\theta\partial\theta'}\log g(y,\theta) \\
\Omega &= E\left(\frac{\partial}{\partial\theta}\log g(y,\theta)\frac{\partial}{\partial\theta}\log g(y,\theta)'\right)
\end{aligned}
$$

If the model is correctly specified ($g(y,\theta_0) = f(y)$), then $Q = \Omega$ (the information matrix equality). Otherwise $Q \neq \Omega$.

15.3 Expected KLIC

The MLE $\hat{\theta} = \hat{\theta}(\mathbf{y})$ is a function of the data vector $\mathbf{y}$.

The fitted model at any $\tilde{\mathbf{y}}$ is $\hat{\mathbf{g}}(\tilde{\mathbf{y}}) = \mathbf{g}(\tilde{\mathbf{y}}, \hat{\theta}(\mathbf{y}))$.

The fitted likelihood is $\mathcal{L}(\hat{\theta}) = -\log \mathbf{g}(\mathbf{y}, \hat{\theta}(\mathbf{y}))$ (the model evaluated at the observed data).

The KLIC of the fitted model is is

$$
\begin{aligned}
KLIC(\mathbf{f}, \hat{\mathbf{g}}) &= C_f - \int \mathbf{f}(\tilde{\mathbf{y}})\log \mathbf{g}(\tilde{\mathbf{y}}, \hat{\theta}(\mathbf{y}))d\tilde{\mathbf{y}} \\
&= C_f - E_{\tilde{\mathbf{y}}}\log \mathbf{g}(\tilde{\mathbf{y}}, \hat{\theta}(\mathbf{y}))
\end{aligned}
$$

where $\tilde{\mathbf{y}}$ has density $\mathbf{f}$, independent of $\mathbf{y}$.

The expected KLIC is the expectation over the observed values $\mathbf{y}$

$$
\begin{aligned}
E\left(KLIC(\mathbf{f}, \hat{\mathbf{g}})\right) &= C_f - E_{\mathbf{y}}E_{\tilde{\mathbf{y}}}\log \mathbf{g}(\tilde{\mathbf{y}}, \hat{\theta}(\mathbf{y})) \\
&= C_f - E_{\tilde{\mathbf{y}}}E_{\mathbf{y}}\log \mathbf{g}(\mathbf{y}, \hat{\theta}(\tilde{\mathbf{y}}))
\end{aligned}
$$

the second equality by symmetry. In this expression, $\tilde{\mathbf{y}}$ and $\mathbf{y}$ are independent vectors each with density $\mathbf{f}$. Letting $\tilde{\theta} = \hat{\theta}(\tilde{\mathbf{y}})$, the estimator of θ when the data is $\tilde{\mathbf{y}}$, we can write this more compactly as

$$
E_{\cdot y}\left(KLIC(\mathbf{f}, \hat{\mathbf{g}})\right) = C_f - E\log \mathbf{g}(\mathbf{y}, \tilde{\theta})
$$

where $\mathbf{y}$ and $\tilde{\theta}$ are independent.

An alternative interpretation is in terms of predicted likelihood. The expected KLIC is the expected likelihood when the sample $\tilde{\mathbf{y}}$ is used to construct the estimate $\tilde{\theta}$, and an independent sample $\mathbf{y}$ used for evaluation. In linear regression, the quasi-likelihood is Gaussian, and the expected KLIC is the expected squared prediction error.

15.4 Estimating KLIC

We want an estimate of the expected KLIC.

As C_f is constant across models, it is ignored.

We want to estimate

$$
T = -E\log \mathbf{g}(\mathbf{y}, \tilde{\theta})
$$

Make a second-order Taylor expansion of $-\log \mathbf{g}\left(\mathbf{y}, \tilde{\theta}\right)$ about $\hat{\theta}$:

$$\begin{aligned} -\log \mathbf{g}(\mathbf{y}, \tilde{\theta}) &\simeq -\log \mathbf{g}(\mathbf{y}, \hat{\theta}) - \frac{\partial}{\partial \theta} \log \mathbf{g}(\mathbf{y}, \hat{\theta})' \left(\tilde{\theta} - \hat{\theta}\right) \\ &\quad -\frac{1}{2}\left(\tilde{\theta} - \hat{\theta}\right)' \left(\frac{\partial^2}{\partial \theta \partial \theta'} \log \mathbf{g}(\mathbf{y}, \hat{\theta})\right)\left(\tilde{\theta} - \hat{\theta}\right) \end{aligned}$$

The first term on the RHS is $\mathcal{L}(\hat{\theta})$, the second is linear in the FOC, so only the third term remains. Writing

$$\hat{Q} = -n^{-1} \frac{\partial^2}{\partial \theta \partial \theta'} \log \mathbf{g}(\mathbf{y}, \hat{\theta}),$$

$$\tilde{\theta} - \hat{\theta} = \left(\tilde{\theta} - \theta_0\right) - \left(\hat{\theta} - \theta_0\right)$$

and expanding the quadratic, we find

$$-\log \mathbf{g}(\mathbf{y}, \tilde{\theta}) \simeq \mathcal{L}(\hat{\theta}) + \frac{1}{2} n \left(\tilde{\theta} - \theta_0\right)' \hat{Q} \left(\tilde{\theta} - \theta_0\right) + \frac{1}{2} n \left(\hat{\theta} - \theta_0\right)' \hat{Q} \left(\hat{\theta} - \theta_0\right) + n \left(\tilde{\theta} - \theta_0\right) \hat{Q} \left(\hat{\theta} - \theta_0\right).$$

Now

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \rightarrow_d Z_1 \sim N(0, V)$$

$$\sqrt{n}\left(\tilde{\theta} - \theta_0\right) \rightarrow_d Z_2 \sim N(0, V)$$

which are independent, and $\hat{Q} \rightarrow_p Q$. Thus for large n,

$$-\log \mathbf{g}(\mathbf{y}, \tilde{\theta}) \simeq \mathcal{L}(\hat{\theta}) + \frac{1}{2} Z_1' Q Z_1 + \frac{1}{2} Z_2' Q Z_2 + Z_1' Q Z_2.$$

Taking expectations

$$\begin{aligned} T &= -E \log \mathbf{g}(\mathbf{y}, \tilde{\theta}) \\ &\simeq E\mathcal{L}(\hat{\theta}) + E\left(\frac{1}{2} Z_1' Q Z_1 + \frac{1}{2} Z_2' Q Z_2 + Z_1' Q Z_2\right) \\ &= E\mathcal{L}(\hat{\theta}) + \operatorname{tr}(QV) \\ &= E\mathcal{L}(\hat{\theta}) + \operatorname{tr}\left(Q^{-1}\Omega\right) \end{aligned}$$

An (asymptotically) unbiased estimate of T is then

$$\hat{T} = \mathcal{L}(\hat{\theta}) + \operatorname{tr}\widehat{(Q^{-1}\Omega)}$$

where $\operatorname{tr}\widehat{(Q^{-1}\Omega)}$ is an estimate of $\operatorname{tr}\left(Q^{-1}\Omega\right)$.

15.5 AIC

When $g(x, \theta_0) = f(x)$ (the model is correctly specified) then $Q = \Omega$ (the information matrix equality). Hence

$$\operatorname{tr}\left(Q^{-1}\Omega\right) = k = \dim(\theta)$$

so

$$\hat{T} = \mathcal{L}(\hat{\theta}) + k$$

This is the the Akaike Information Criterion (AIC). It is typically written as $2\hat{T}$, e.g.

$$AIC = 2\mathcal{L}(\hat{\theta}) + 2k$$

AIC is an estimate of the expected KLIC, based on the approximation that g includes the correct model.

Picking a model with the smalled AIC is picking the model with the smallest estimated KLIC. In this sense it is picking is the best-fitting model.

15.6 TIC

Takeuchi (1976) proposed a robust AIC, and is known as the Takeuchi Information Criterion (TIC)

$$TIC = 2\mathcal{L}(\hat{\theta}) + 2\operatorname{tr}\left(\hat{Q}^{-1}\hat{\Omega}\right)$$

where

$$\begin{aligned}
\hat{Q} &= -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}\log g(y_i, \hat{\theta}) \\
\hat{\Omega} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\log g(y_i, \hat{\theta})\frac{\partial}{\partial\theta}\log g(y_i, \hat{\theta})'\right)
\end{aligned}$$

The does not require that g is correctly specified.

15.7 Comments on AIC and TIC

The AIC and TIC are designed for the likelihood (or quasi-likelihood) context. For proper application, the "model" needs to be a conditional density, not just a conditional mean or set of moment conditions. This is a strength and limitation.

The benefit of AIC/TIC is that it selects fitted models whose densities are close to the true density. This is a broad and useful feature.

The relation of the TIC to the AIC is very similar to the relationship between the conventional and "White" covariance matrix estimators for the MLE/QMLE or LS. The TIC does not appear to be widely appreciated nor used.

The AIC is known to be asymptotically optimal in linear regression (we discuss this below), but in the general context I do not know of an optimality result. The desired optimality would be that if a model is selected by minimizing AIC (or TIC) then the fitted KLIC of this model is asymptotically equivalent to the KLIC of the infeasible best-fitting model.

15.8 AIC and TIC in Linear Regression

In linear regression or projection

$$\begin{aligned} y_i &= X_i'\theta + e_i \\ E\left(X_i e_i\right) &= 0 \end{aligned}$$

AIC or TIC cannot be directly applied, as the density of e_i is unspecified. However, the LS estimator is the same as the Gaussian MLE, so it is natural to calculate the AIC or TIC for the Gaussian quasi-MLE.

The Gaussian quasi-likelihood is

$$\log g_i(\theta) = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(y_i - X_i'\beta\right)^2$$

where $\theta = (\beta, \sigma^2)$ and $\sigma^2 = Ee_i^2$. The MLE $\hat{\theta} = (\hat{\beta}, \hat{\sigma}^2)$ is LS. The pseudo-true value β_0 is the projection coefficient $\beta = E\left(X_i X_i'\right)^{-1} E\left(X_i y_i\right)$. If β is $k \times 1$ then the number of parameters is $k+1$.

The sample log-likelihood is

$$2\mathcal{L}(\hat{\theta}) = n\log\left(\hat{\sigma}^2\right) + n\log\left(2\pi\right) + n$$

The second/third parts can be ignored. The AIC is

$$AIC = n\log\left(\hat{\sigma}^2\right) + 2\left(k+1\right).$$

Often this is written

$$AIC = n\log\left(\hat{\sigma}^2\right) + 2k$$

as adding/subtracting constants do not matter for model selection, or sometimes

$$AIC = \log\left(\hat{\sigma}^2\right) + 2\frac{k}{n}$$

as scaling doesn't matter.

Also

$$\begin{aligned}
\frac{\partial}{\partial \beta} \log g(y_i, \theta) &= \frac{1}{\sigma^2} X_i \left(y_i - X_i' \beta \right) \\
\frac{\partial}{\partial \sigma^2} \log g(y_i, \theta) &= -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} \left(y_i - X_i' \beta \right)^2,
\end{aligned}$$

and

$$\begin{aligned}
-\frac{\partial^2}{\partial \beta \partial \beta'} \log g(y_i, \theta) &= \frac{1}{\sigma^2} X_i X_i' \\
-\frac{\partial^2}{\partial \beta \partial \sigma^2} \log g(y_i, \theta) &= \frac{1}{\sigma^4} X_i \left(y_i - X_i' \beta \right) \\
-\frac{\partial^2}{\partial \left(\sigma^2\right)^2} \log g(y_i, \theta) &= -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6} \left(y_i - X_i' \beta \right)^2
\end{aligned}$$

Evaluated at the pseudo-true values,

$$\begin{aligned}
\frac{\partial}{\partial \beta} \log g(y_i, \theta_0) &= \frac{1}{\sigma^2} X_i e_i \\
\frac{\partial}{\partial \sigma^2} \log g(y_i, \theta_0) &= \frac{1}{2\sigma^4} \left(e_i^2 - \sigma^2 \right),
\end{aligned}$$

and

$$\begin{aligned}
-\frac{\partial^2}{\partial \beta \partial \beta'} \log g(y_i, \theta_0) &= \frac{1}{\sigma^2} X_i X_i' \\
-\frac{\partial^2}{\partial \beta \partial \sigma^2} \log g(y_i, \theta_0) &= \frac{1}{\sigma^4} X_i e \\
-\frac{\partial^2}{\partial \left(\sigma^2\right)^2} \log g(y_i, \theta_0) &= \frac{1}{2\sigma^6} \left(2 \left(y_i - X_i' \beta \right)^2 - \sigma^2 \right)
\end{aligned}$$

Thus

$$\begin{aligned}
Q &= -E \begin{bmatrix} \frac{\partial^2}{\partial \beta \partial \beta'} \log g(y_i, \theta_0) & \frac{\partial^2}{\partial \beta \partial \sigma^2} \log g(y_i, \theta_0) \\ \frac{\partial^2}{\partial \sigma^2 \partial \beta'} \log g(y_i, \theta_0) & \frac{\partial^2}{\partial \left(\sigma^2\right)^2} \log g(y_i, \theta_0) \end{bmatrix} \\
&= \sigma^{-2} \begin{bmatrix} E \left(X_i X_i' \right) & 0 \\ 0 & \frac{1}{2\sigma^2} \end{bmatrix}
\end{aligned}$$

and

$$\begin{aligned}\Omega &= E\left[\begin{array}{cc}\frac{\partial}{\partial\beta}\log g(y_i,\theta_0)\frac{\partial}{\partial\beta}\log g(y_i,\theta_0)' & \frac{\partial}{\partial\beta}\log g(y_i,\theta_0)\frac{\partial}{\partial\sigma^2}\log g(y_i,\theta_0) \\ \frac{\partial}{\partial\sigma^2}\log g(y_i,\theta_0)\frac{\partial}{\partial\beta}\log g(y_i,\theta_0)' & \left(\frac{\partial}{\partial\sigma^2}\log g(y_i,\theta_0)\right)^2\end{array}\right] \\ &= \sigma^{-2}\left[\begin{array}{cc}E\left(X_iX_i'\frac{e_i^2}{\sigma^2}\right) & \frac{1}{2\sigma^4}E\left(X_i'e_i^3\right) \\ \frac{1}{2\sigma^4}E\left(X_ie_i^3\right) & \frac{\kappa_4}{4\sigma^2}\end{array}\right]\end{aligned}$$

where

$$\kappa_4 = var\left(\frac{e_i^2}{\sigma^2}\right) = \frac{E\left(e_i^2-\sigma^2\right)^2}{\sigma^4} = \frac{E\left(e_i^4\right)-\sigma^4}{\sigma^4}$$

We see that $\Omega = Q$ if

$$\begin{aligned}E\left(\frac{e_i^2}{\sigma^2} \mid X_i\right) &= 1 \\ E\left(X_ie_i^3\right) &= 0 \\ \kappa_4 &= 2\end{aligned}$$

Essentially, this requies that $e_i \sim N(0,\sigma^2)$. Otherwise $\Omega \neq Q$.

Thus the AIC is appropriate in Gaussian regression. It is an “approximation” in non-Gaussian regression, heteroskedastic regression, or projection.

To calculate the TIC, note that since Q is block diagonal you do not need to estimate the off-diagonal component of Ω. Note that

$$\begin{aligned}\operatorname{tr} Q^{-1}\Omega &= \operatorname{tr}\left[E\left(X_iX_i'\right)^{-1}E\left(X_iX_i'\frac{e_i^2}{\sigma^2}\right)\right] + \left(\frac{1}{2\sigma^2}\right)^{-1}\frac{\kappa_4}{4\sigma^2} \\ &= \operatorname{tr}\left[E\left(X_iX_i'\right)^{-1}E\left(X_iX_i'\frac{e_i^2}{\sigma^2}\right)\right] + \frac{\kappa_4}{2}\end{aligned}$$

Let

$$\hat{\kappa}_4 = \frac{1}{n}\sum_{i=1}^n\left(\hat{e}_i^2-\hat{\sigma}^2\right)^2$$

The TIC is then

$$\begin{aligned}TIC &= n\log\left(\hat{\sigma}^2\right) + \operatorname{tr}\left(\hat{Q}^{-1}\hat{\Omega}\right) \\ &= n\log\left(\hat{\sigma}^2\right) + 2\left[\operatorname{tr}\left(\left(\sum_{i=1}^n X_iX_i'\right)^{-1}\left(\sum_{i=1}^n X_iX_i'\frac{\hat{e}_i^2}{\hat{\sigma}^2}\right)\right) + \hat{\kappa}_4\right] \\ &= n\log\left(\hat{\sigma}^2\right) + \frac{2}{\hat{\sigma}^2}\sum_{i=1}^n h_i\hat{e}_i^2 + \hat{\kappa}_4\end{aligned}$$

where $h_i = X_i' \left(\mathbf{X}'\mathbf{X}\right)^{-1} X_i$.

When the errors are close to homoskedastic and Gaussian, then h_i and e_i^2 will be uncorrelated $\hat{\kappa}_4$ will be close to 2, so the penalty will be close to

$$2\sum_{i=1}^{n} h_i + 2 = 2\left(k+1\right)$$

as for AIC. In this case TIC will be close to AIC. In applications, the differences will arise under heteroskedasticity and non-Gaussianity.

The primary use of AIC and TIC is to compare models. As we change models, typically the residuals $\hat{e}_i$ do not change too much, so my guess is that the estimate $\hat{\kappa}$ will not change much. In this event, the TIC correction for estimation of σ^2 will not matter much.

15.9 Asymptotic Equivalence

Let $\tilde{\sigma}^2$ be a preliminary (model-free) estimate of σ^2. The AIC is equivalent to

$$\begin{aligned}
\tilde{\sigma}^2\left(AIC - n\log\tilde{\sigma}^2 + n\right) &= n\sigma^2\left(\log\left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right) + 1\right) + 2\tilde{\sigma}^2 k \\
&\simeq n\tilde{\sigma}^2\left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right) + 2\tilde{\sigma}^2 k \\
&= \hat{e}'\hat{e} + 2\tilde{\sigma}^2 k \\
&= C_k
\end{aligned}$$

The approximation is $\log(1+a) \simeq a$ for a small. This is the Mallows criterion. Thus AIC is approximately equal to Mallows, and the approximation is close when k/n is small.

Furtheremore, this expression approximately equalts

$$\hat{e}'\hat{e}\left(1 + \frac{2}{nk}\right) = S_k$$

which is known as Shibata's condition (Annals of Statistics, 1980; Biometrick, 1981).

The TIC (ignoring the correction for estimation of σ^2) is equivalent to

$$\begin{aligned}
\tilde{\sigma}^2\left(TIC - n\log\tilde{\sigma}^2 + n\right) &= n\tilde{\sigma}^2\left(\log\left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right) + 1\right) + \frac{2\tilde{\sigma}^2}{\hat{\sigma}^2}\sum_{i=1}^{n} h_i\hat{e}_i^2 \\
&\simeq \hat{e}'\hat{e} + 2\sum_{i=1}^{n} h_i\hat{e}_i^2 \\
&\simeq \sum_{i=1}^{n}\frac{\hat{e}_i^2}{\left(1-h_i\right)^2} \\
&= CV,
\end{aligned}$$

the cross-validation criterion. Thus $TIC \simeq CV$.

They are both asymptotically equivalent to a "Heteroskedastic-Robust Mallows Criterion"

$$C_k^* = \hat{e}'\hat{e} + 2\sum_{i=1}^{n} h_i \hat{e}_i^2$$

which, strangely enough, I have not seen in the literature.

15.10 Mallows Criterion

Ker-Chau Li (1987, Annals of Statistics) provided a important treatment of the optimality of model selection methods for homoskedastic linear regression. Andrews (1991, JoE) extended his results to allow conditional heteroskedasticity.

Take the regression model

$$\begin{aligned} y_i &= g\left(X_i\right) + e_i \\ &= g_i + e_i \\ E\left(e_i \mid X_i\right) &= 0 \\ E\left(e_i^2 \mid X_i\right) &= 0 \end{aligned}$$

Written as an $n \times 1$ vector

$$y = g + e.$$

Li assumed that the X_i are non-random, but his analysis can be re-interpreted by treating everything as conditional on X_i.

Li considered estimators of the $n \times 1$ vector g which are linear in y and thus take the form

$$\hat{g}(h) = M(h)y$$

where $M(h)$ is $n \times n$, a function of the X matrix, indexed by $h \in H$, and H is a discrete set. For example, a series estimator sets $M(h) = X_h \left(X_h' X_h\right)^{-1} X_h$ where X_h is an $n \times k_h$ set of basis functions of the regressors, and $H = \{1, ..., \bar{h}\}$. The goal is to pick h to minimize the average squared error

$$L(h) = \frac{1}{n}\left(g - \hat{g}(h)\right)'\left(g - \hat{g}(h)\right).$$

The index h is selected by minimizing the Mallows, Generalized CV, or CV criterion. We discuss Mallows in detail, as it is the easiest to analyze. Andrews showed that only CV is optimal under heteroskedasticity.

The Mallows criterion is

$$C(h) = \frac{1}{n}\left(y - \hat{g}(h)\right)'\left(y - \hat{g}(h)\right) + \frac{2\sigma^2}{n} \operatorname{tr} M(h)$$

The first term is the residual variance from model h, the second is the penalty. For series estimators, $\operatorname{tr} M(h) = k_h$. The Mallows selected index $\hat{h}$ minimizes $C(h)$.

Since $y = e + g$, then $y - \hat{g}(h) = e + g - \hat{g}(h)$, so

$$\begin{aligned} C(h) &= \frac{1}{n}(y - \hat{g}(h))'(y - \hat{g}(h)) + \frac{2\sigma^2}{n}\operatorname{tr} M(h) \\ &= \frac{1}{n}e'e + L(h) + 2\frac{1}{n}e'(g - \hat{g}(h)) + \frac{2\sigma^2}{n}\operatorname{tr} M(h) \end{aligned}$$

And

$$\hat{g}(h) = M(h)y = M(h)g + M(h)e$$

then

$$\begin{aligned} g - \hat{g}(h) &= (I - M(h))\, g - M(h)e \\ &= b(h) - M(h)e \end{aligned}$$

where $b(h) = (I - M(h))\, g$, and $C(h)$ equals

$$\frac{1}{n}e'e + L(h) + 2\frac{1}{n}e'b(h) + \frac{2}{n}\left(\sigma^2 \operatorname{tr} M(h) - e'M(h)e\right)$$

As the first term doesn't involve h, it follows that $\hat{h}$ minimizes

$$C^*(h) = L(h) + 2\frac{1}{n}e'b(h) + \frac{2}{n}\left(\sigma^2 \operatorname{tr} M(h) - e'M(h)e\right)$$

over $h \in H$.

The idea is that empirical criterion $C^*(h)$ equals the desired criterion $L(h)$ plus a stochastically small error.

We calculate that

$$\begin{aligned} L(h) &= \frac{1}{n}(g - \hat{g}(h))'(g - \hat{g}(h)) \\ &= \frac{1}{n}b(h)'b(h) - \frac{2}{n}b(h)'M(h)e + \frac{1}{n}e'M(h)'M(h)e \end{aligned}$$

and

$$\begin{aligned} R(h) &= E\left(L(h) \mid \mathbf{X}\right) \\ &= \frac{1}{n}b(h)'b(h) + E\left(\frac{1}{n}e'M(h)'M(h)e \mid \mathbf{X}\right) \\ &= \frac{1}{n}b(h)'b(h) + \frac{\sigma^2}{n}\operatorname{tr} M(h)'M(h) \end{aligned}$$

The optimality result is:

Theorem 1. Let $\lambda_{\max}(A)$ denote the maximum eigenvalue of A. If for some positive integer m,

$$\begin{aligned} \lim_{n\to\infty} \sup_{h\in H} \lambda_{\max}\left(M(h)\right) &< \infty \\ E\left(e_i^{4m} \mid X_i\right) &\leq \kappa < \infty \\ \sum_{h\in H} \left(nR(h)\right)^{-m} &\to 0 \end{aligned} \tag{1}$$

then

$$\frac{L(\hat{h})}{\inf_{h\in H} L(h)} \to_p 1.$$

15.11 Whittle's Inequalities

To prove Theorem 1, Li (1987) used two key inequalities from Whittle (1960, Theory of Probability and Its Applications).

Theorem. Suppose the observations are independent. Let $\mathbf{b}$ be any $n \times 1$ vector and $\mathbf{A}$ any $n \times n$ matrix, functions of $\mathbf{X}$. If for some $s \geq 2$

$$\max_i E\left(|e_i|^s \mid X_i\right) \leq \kappa_s < \infty$$

then

$$E\left(\left|\mathbf{b}'\mathbf{e}\right|^s \mid \mathbf{X}\right) \leq K_{1s}\left(\mathbf{b}'\mathbf{b}\right)^{s/2} \tag{2}$$

and

$$E\left(\left|\mathbf{e}'\mathbf{A}\mathbf{e} - E\left(\mathbf{e}'\mathbf{A}\mathbf{e} \mid \mathbf{X}\right)\right|^s \mid \mathbf{X}\right) \leq K_{2s}\left(\operatorname{tr}\mathbf{A}'\mathbf{A}\right)^{s/2} \tag{3}$$

where

$$K_{1s} = \frac{2^{s3/2}}{\sqrt{\pi}}\Gamma\left(\frac{s+1}{2}\right)\kappa_s$$

$$K_{2s} = 2^s K_{1s} K_{1,2s}^{1/2} \kappa_{2s}$$

15.12 Proof of Theorem 1

The main idea is similar to that of consistent estimation. Recall that if $S_n(\theta) \to_p S(\theta)$ uniformly in θ, then the minimizer of $S_n(\theta)$ converges to the minimizer of $S(\theta)$. We can write the uniform convergence as

$$\sup_\theta \left|\frac{S_n(\theta)}{S(\theta)} - 1\right| \to_p 0$$

In the present case, we will show (below) that

$$\sup_h \left|\frac{C^*(h) - L(h)}{L(h)}\right| \to_p 0 \tag{4}$$

Let h_0 denote the minimizer of $L(h)$. Then

$$\begin{aligned}
0 &\leq \frac{L(\hat{h}) - L(h_0)}{L(\hat{h})} \\
&= \frac{C^*(\hat{h}) - L(h_0)}{L(\hat{h})} - \frac{C^*(\hat{h}) - L(\hat{h})}{L(\hat{h})} \\
&= \frac{C^*(\hat{h}) - L(h_0)}{L(\hat{h})} + o_p(1) \\
&\leq \frac{C^*(h_0) - L(h_0)}{L(\hat{h})} + o_p(1) \\
&\leq \frac{C^*(h_0) - L(h_0)}{L(h_0)} + o_p(1) \\
&= o_p(1)
\end{aligned}$$

This uses (4) twice, and the facts $L(h_0) \leq L(\hat{h})$ and $C^*(\hat{h}) \leq C^*(h_0)$. This shows that

$$\frac{L(h_0)}{L(\hat{h})} \rightarrow_p 1$$

which is equivalent to the Theorem.

The key is thus (4). We show below that

$$\sup_h \left| \frac{L(h)}{R(h)} - 1 \right| \rightarrow_p 0 \tag{5}$$

which says that $L(h)$ and $R(h)$ are asymptotically equivalent, and thus (4) is equivalent to

$$\sup_h \left| \frac{C^*(h) - L(h)}{R(h)} \right| \rightarrow_p 0. \tag{6}$$

From our earlier equation for $C^*(h)$, we have

$$\sup_h \left| \frac{C^*(h) - L(h)}{R(h)} \right| \leq 2 \sup_h \frac{|e'b(h)|}{nR(h)} + 2 \sup_h \frac{\left|\sigma^2 \operatorname{tr} M(h) - e'M(h)e\right|}{nR(h)}. \tag{7}$$

Take the first term on the right-hand-side. By Whittle's first inequality,

$$E\left(\left| e'b(h) \right|^{2m} \mid \mathbf{X} \right) \leq K \left(b(h)'b(h) \right)^m$$

Now recall

$$nR(h) = b(h)'b(h) + \sigma^2 \operatorname{tr} M(h)'M(h) \tag{8}$$

Thus

$$nR(h) \geq b(h)'b(h)$$

Hence

$$E\left(\left|e'b(h)\right|^{2m} \mid \mathbf{X}\right) \leq K\left(b(h)'b(h)\right)^{m} \leq K\left(nR(h)\right)^{m}$$

Then, since H is discrete, by applying Markov's inequality and this bound,

$$\begin{aligned} P\left(\sup_{h} \frac{\left|e'b\left(h\right)\right|}{nR(h)} > \delta \mid \mathbf{X}\right) &\leq \sum_{h\in H} P\left(\frac{\left|e'b(h)\right|}{nR(h)} > \delta \mid \mathbf{X}\right) \\ &\leq \sum_{h\in H} \delta^{-2m} \frac{E\left(\left|e'b(h)\right|^{2m} \mid \mathbf{X}\right)}{\left(nR(h)\right)^{2m}} \\ &\leq \sum_{h\in H} \delta^{-2m} \frac{K\left(nR(h)\right)^{m}}{\left(nR(h)\right)^{2m}} \\ &= \frac{K}{\delta^{2m}} \sum_{h\in H} \left(nR(h)\right)^{-m} \\ &\to 0 \end{aligned}$$

by assumption (1). This shows

$$\sup_{h} \frac{\left|e'b(h)\right|}{nR(h)} \to_{p} 0$$

Now take the second term in (7). By Whittle's second inequality, since

$$E\left(e'M(h)e \mid \mathbf{X}\right) = \sigma^{2} \operatorname{tr} M(h),$$

then

$$\begin{aligned} E\left(\left|e'M(h)e - \sigma^{2} \operatorname{tr} M(h)\right|^{2m} \mid \mathbf{X}\right) &\leq K\left(\operatorname{tr}\left(M(h)'M(h)\right)\right)^{m} \\ &\leq \sigma^{-2m} K\left(nR(h)\right)^{m} \end{aligned}$$

the second inequality since (8) implies

$$\operatorname{tr} M(h)'M(h) \leq \sigma^{-2} nR(h)$$

Applying Markov's inequality

$$\begin{aligned}
P\left(\sup_h \frac{\left|e'M(h)e-\sigma^2\operatorname{tr}M(h)\right|}{nR(h)} > \delta \mid \mathbf{X}\right) &\le \sum_{h\in H} P\left(\frac{\left|e'M(h)e-\sigma^2\operatorname{tr}M(h)\right|}{nR(h)} > \delta \mid \mathbf{X}\right) \\
&\le \sum_{h\in H}\delta^{-2m}\frac{E\left(\left|e'M(h)e-\sigma^2\operatorname{tr}M(h)\right|^{2m} \mid \mathbf{X}\right)}{\left(nR(h)\right)^{2m}} \\
&\le \sum_{h\in H}\delta^{-2m}\frac{\sigma^{-2m}K\left(nR(h)\right)^m}{\left(nR(h)\right)^{2m}} \\
&= K\left(\delta^2\sigma^2\right)^{-m}\sum_{h\in H}\left(nR(h)\right)^{-m} \\
&\to 0
\end{aligned}$$

For completeness, let us show (5). The demonstration is essentially the same as the above. We calculate

$$\begin{aligned}
L(h)-R(h) &= -\frac{2}{n}b(h)'M(h)e+\frac{1}{n}e'M(h)'M(h)e-\frac{\sigma^2}{n}\operatorname{tr}M(h)'M(h) \\
&= -\frac{2}{n}b(h)'M(h)e+\frac{1}{n}\left(e'M(h)'M(h)e-E\left(e'M(h)'M(h)e \mid \mathbf{X}\right)\right)
\end{aligned}$$

Thus

$$\sup_h\left|\frac{L(h)-R(h)}{R(h)}\right| \le 2\sup_h\frac{|e'M(h)'b(h)|}{nR(h)}+2\sup_h\frac{|e'M(h)'M(h)e-E\left(e'M(h)'M(h)e \mid \mathbf{X}\right)|}{nR(h)}.$$

By Whittle's first inequality,

$$E\left(\left|e'M(h)'b(h)\right|^{2m} \mid \mathbf{X}\right) \le K\left(b(h)'M(h)M(h)'b(h)\right)^m$$

Use the matrix inequality

$$\operatorname{tr}(AB) \le \lambda_{\max}(A)\operatorname{tr}(B)$$

and letting

$$\bar{M}=\lim_{n\to\infty}\sup_{h\in H}\lambda_{\max}\left(M(h)\right)<\infty$$

then

$$\begin{aligned}
b(h)'M(h)M(h)'b(h) &= \operatorname{tr}\left(M(h)M(h)'b(h)b(h)'\right) \\
&\le \bar{M}^2\operatorname{tr}\left(b(h)b(h)'\right) \\
&\le \bar{M}^2 b(h)'b(h) \\
&\le \bar{M}^2 nR(h)
\end{aligned}$$

Thus

$$\begin{aligned} E\left(\left|e'M(h)'b(h)\right|^{2m} \mid \mathbf{X}\right) &\leq K\left(b(h)'M(h)M(h)'b(h)\right)^m \\ &\leq K\bar{M}^2\left(nR(h)\right)^m \end{aligned}$$

Thus

$$\begin{aligned} P\left(\sup_h \frac{|e'M(h)'b\,(h)|}{nR(h)} > \delta \mid \mathbf{X}\right) &\leq \sum_{h\in H} P\left(\frac{|e'M(h)'b(h)|}{nR(h)} > \delta \mid \mathbf{X}\right) \\ &\leq \sum_{h\in H} \delta^{-2m}\frac{E\left(|e'M(h)'b(h)|^{2m} \mid \mathbf{X}\right)}{\left(nR(h)\right)^{2m}} \\ &\leq \sum_{h\in H} \delta^{-2m}\frac{K\bar{M}^2\left(nR(h)\right)^m}{\left(nR(h)\right)^{2m}} \\ &= \frac{K\bar{M}^2}{\delta^{2m}}\sum_{h\in H}\left(nR(h)\right)^{-m} \\ &\to 0 \end{aligned}$$

Similarly,

$$\begin{aligned} E\left(\left|e'M(h)'M(h)e - E\left(e'M(h)'M(h)e \mid \mathbf{X}\right)\right|^{2m} \mid \mathbf{X}\right) &\leq K\left(\operatorname{tr}\left(M(h)'M(h)M(h)'M(h)\right)\right)^m \\ &\leq K\bar{M}^{2m}\left(\operatorname{tr}\left(M(h)'M(h)\right)\right)^m \\ &\leq \sigma^{-2m}K\bar{M}^{2m}\left(nR(h)\right)^m \end{aligned}$$

and thus

$$\begin{aligned} P\left(\sup_h \frac{|e'M(h)'M(h)e - E\left(e'M(h)'M(h)e \mid \mathbf{X}\right)|}{nR(h)} > \delta \mid \mathbf{X}\right) &\leq \sum_{h\in H} P\left(\frac{|e'M(h)'M(h)e - E\left(e'M(h)'M(h)e \mid \mathbf{X}\right.}{nR(h)}\right. \\ &\leq \sum_{h\in H} \delta^{-2m}\frac{E\left(|e'M(h)'M(h)e - E\left(e'M(h)'M(h)e\right.\right.}{\left(nR(h)\right)^{2m}} \\ &\leq \sum_{h\in H} \delta^{-2m}\frac{\sigma^{-2m}K\bar{M}^{2m}\left(nR(h)\right)^m}{\left(nR(h)\right)^{2m}} \\ &= K\left(\frac{\bar{M}^2}{\delta^2\sigma^2}\right)^m\sum_{h\in H}\left(nR(h)\right)^{-m} \\ &\to 0 \end{aligned}$$

We have shown

$$\sup_h\left|\frac{L(h)-R(h)}{R(h)}\right| \to_p 0$$

which is (5).

15.13 Mallows Model Selection

Li's Theorem 1 applies to a variety of linear estimators. Of particular interest is model selection (e.g. series estimation).

Let's verify Li's conditions, which were

$$\begin{aligned} \lim_{n\to\infty} \sup_{h\in H} \lambda_{\max}\left(M(h)\right) &< \infty \\ E\left(e_i^{4m} \mid X_i\right) &\leq \kappa < \infty \\ \sum_{h\in H} \left(nR(h)\right)^{-m} &\to 0 \end{aligned} \tag{9}$$

In linear estimation, $M(h)$ is a projection matrix, so $\lambda_{\max}\left(M(h)\right) = 1$ and the first equation is automatically satisfied.

The key is equation (9).

Suppose that for sample size n, there are N_n models. Let

$$\xi_n = \inf_{h\in H} nR(h)$$

and assume

$$\xi_n \to \infty$$

A crude bound is

$$\sum_{h\in H} \left(nR(h)\right)^{-m} \leq N_n \xi_n^{-m}$$

If $N_n \xi_n^{-m} \to 0$ then (9) holds. Notice that by increasing m, we can allow for larger N_n (more models) but a tighter moment bound.

The condition $\xi_n \to 0$ says that for all finite models h, the there is non-zero approximation error, so that $R(h)$ is non-zero. In contrast, if there is a finite dimensional model h_0 for which $b(h_0) = 0$, then $nR(h_0) = h_0\sigma^2$ does not diverge. In this case, Mallows (and AIC) are asymptotically sub-optimal.

We can improve this condition if we consider the case of selection among models of increasing size. Suppose that model h has k_h regressors, and $k_1 < k_2 < \cdots$ and for some $m \geq 2$,

$$\sum_{h=1}^{\infty} k_h^{-m} < \infty$$

This includes nested model selection, where $k_h = h$ and $m = 2$. Note that

$$nR(h) = b(h)'b(h) + k_h\sigma^2 \geq k_h\sigma^2$$

Now pick $B_n \to \infty$ so that $B_n \xi_n^{-m} \to 0$ (which is possible since $\xi_n \to \infty$.) Then

$$\begin{aligned}\sum_{h=1}^{\infty} (nR(h))^{-m} &= \sum_{h=1}^{B_n} (nR(h))^{-m} + \sum_{h=B_n+1}^{\infty} (nR(h))^{-m} \\ &\leq B_n \xi_n^{-m} + \sigma^{-2} \sum_{h=B_n+1}^{\infty} k_h^{-m} \to 0\end{aligned}$$

as required.

15.14 GMM Model Selection

This is an underdeveloped area. I list a few papers.

Andrews (1999, Econometrica). He considers selecting moment conditions to be used for GMM estimation. Let p be the number of parameters, c represent a list of "selected" moment conditions, $|c|$ denote the cardinality (number) of these moments, and $J_n(c)$ the GMM criterion computed using these c moments. Andrews' proposes criteria of the form

$$IC(c) = J_n(c) - r_n (|c| - p)$$

where $|c| - p$ is the number of overidentifying restrictions and r_n is a sequence. For an AIC-like criterion, he sets $r_n = 2$, for a BIC-like criterion, he sets $r_n = \log n$.

The model selection rule picks the moment conditions c which minimize $J_n(c)$.

Assuming that a subset of the moments are incorrect, Andrews shows that the BIC-like rule asymptotically selects the correct subset.

Andrews and Lu (2001, JoE) extend the above analysis to the case of jointly picking the moments and the parameter vector (that is, imposing zero restrictions on the parameters). They show that the same criterion has similar properties – that it can asymptotically select the "correct" moments and "correct" zero restrictions.

Hong, Preston and Shum (ET, 2003) extend the analysis of the above papers to empirical likelihood. They show that that this criterion has the same interpretation when $J_n(c)$ is replaced by the empirical likelihood.

These papers are an interesting first step, but they do not address the issue of GMM selection when the true model is potentially infinite dimensional and/or misspecified. That is, the analysis is not analogous to that of Li (1987) for the regression model.

In order to properly understand GMM selection, I believe we need to understand the behavior of GMM under misspecification.

Hall and Inoue (2003, Joe) is one of the few contributions on GMM under misspecificaiton. They did not investigate model selection.

Suppose that the model is

$$m(\theta) = Em_i(\theta) = 0$$

where m_i is $\ell \times 1$ and θ is $k \times 1$. Assume $\ell > r$ (overidentification). The model is misspecified if there is no θ such that this moment condition holds. That is, for all θ,

$$m(\theta) \neq 0$$

Suppose we apply GMM. What happens?

The first question is, what is the pseudo-true value? The GMM criterion is

$$J_n(\theta) = n\bar{m}_n(\theta)'W_n\bar{m}_n(\theta)$$

If $W_n \to_p W$, then

$$n^{-1}J_n(\theta) \to_p m(\theta)'Wm(\theta).$$

Thus the GMM estimator $\hat{\theta}$ is consistent for the pseudo-true value

$$\theta_0(W) = \operatorname{argmin} m(\theta)'Wm(\theta).$$

Interstingly, the pseudo-true value $\theta_0(W)$ is a function of W. This is a fundamental difference from the correctly specified case, where the weight matrix only affects efficiency. In the misspecified case, it affects what is being estimated.

This means that when we apply "iterated GMM", the pseudo-true value changes with each step of the iteration!

Hall and Inoue also derive the distribution of the GMM estimator. They find that the distribution depends not only on the randomness in the moment conditions, but on the randomness in the weight matrix. Specifically, they assume that $n^{1/2}(W_n - W) \to_d Normal$, and find that this affects the asymptotic distributions.

Furthermore, the distribution of test statistics is non-standard (a mixture of chi-squares). So inference on the pseudo-true values is troubling.

This subject deserves more study.

15.15 KLIC for Moment Condition Models Under Misspecification

Suppose that the true density is $f(y)$, and we have an over-identified moment condition model, e.g. for some function $m(y)$, the model is

$$Em(y) = 0$$

However, we want to allow for misspecification, namely that

$$Em(y) \neq 0$$

To explore misspefication, we have to ask: What is a desirable pseudo-true model?

Temporarily ignoring parameter estimation, we can ask: Which density $g(y)$ satisfying this moment condition is closest to $f(y)$ in the sense of minimizing KLIC? We can call this $g_0(y)$ the pseudo-true density.

The solution is nicely explained in Appendix A of Chen, Hong, and Shum (JoE, 2007). Recall

$$KLIC(f,g) = \int f(y) \log \left(\frac{f(y)}{g(y)} \right) dy$$

The problem is

$$\min_g KLIC(f,g)$$

subject to

$$\begin{aligned} \int g(y) dy &= 1 \\ \int m(y) g(y) dy &= 0 \end{aligned}$$

The Lagrangian is

$$\int f(y) \log \left(\frac{f(y)}{g(y)} \right) dy + \mu \left(\int g(y) dy - 1 \right) + \lambda' \int m(y) g(y) dy$$

The FOC with respect to $g(y)$ at some y is

$$0 = -\frac{f(y)}{g(y)} + \mu + \lambda' m(y)$$

Multiplying by $g(y)$ and integrating,

$$\begin{aligned} 0 &= -\int f(y) dy + \mu \int g(y) dy + \lambda' \int m(y) g(y) dy \\ &= -1 + \mu \end{aligned}$$

so $\mu = 1$. Solving for $g(y)$ we find

$$g(y) = \frac{f(y)}{1 + \lambda' m(y)},$$

a tilted version of the true density $f(y)$. Inserting this solution we find

$$KLIC(f,g) = \int f(y) \log \left(1 + \lambda' m(y) \right) dy$$

By duality, the optimal Lagrange multiplier λ_0 maximizes this expression

$$\lambda_0 = \underset{\lambda}{\text{argmax}} \int f(y) \log \left(1 + \lambda' m(y) \right) dy.$$

The pseudo-true density is

$$g_0(y) = \frac{f(y)}{1 + \lambda_0' m(y)},$$

with associated minimized KLIC

$$\begin{aligned} KLIC\,(f, g_0) &= \int f(y) \log \left(1 + \lambda_0' m(y)\right) dy \\ &= E \log \left(1 + \lambda_0' m(y)\right) \end{aligned}$$

This is the smallest possible KLIC(f,g) for moment condition models.

This solution looks like empirical likelihood. Indeed, EL minimizes the empirical KLIC, and this connection is widely used to motivate EL.

When the moment $m(y, \theta)$ depends on a parameter θ,then the pseudo-true values (θ_0, λ_0) are the joint solution to the problem

$$\min_\theta \max_\lambda E \log \left(1 + \lambda' m(y, \theta)\right)$$

Theorem (Chen, Hong and Shum, JoE, 2007). If $|m(y, \theta)|$ is bounded, then the EL estimates $(\hat{\theta}, \hat{\lambda})$ are $n^{-1/2}$ consistent for the pseudo-true values (θ_0, λ_0).

This gives a simple interpretation to the definition of KLIC under misspecification.

15.16 Schennach's Impossibility Result

Schennach (Annals of Statistics, 2007) claims a fundamental flaw in the application of KLIC to moment condition models. She shows that the assumption of bounded $|m(y, \theta)|$ is not merely a technical condition, it is binding.

[Notice: In the linear model, $m(y, \theta) = z(y - x'\theta)$ is unbounded if the data has unbounded support. Thus the assumption is highly relevant.]

The key problem is that for any $\lambda \neq 0$, if $m(y, \theta)$ is unbounded, so is $1 + \lambda' m(y, \theta)$. In particular, it can take on negative values. Thus $\log \left(1 + \lambda' m(y, \theta)\right)$ is ill-defined. Thus there is no pseudo-true value of λ. (It must be non-zero, but it cannot be non-zero!) Without a non-zero λ, there is no way to define a pseudo-true θ_0 which satisfies the moment condition.

Technically, Schennach shows that when there is no θ such that $Em(y, \theta) = 0$ and $m(y, \theta)$ is unbounded, then there is no θ_0 such that $\sqrt{n}\left(\hat{\theta} - \theta_0\right) = O_p(1)$.

Her paper leaves open the question: For what is $\hat{\theta}$ consistent? Is there a pseudo-true value? One possibility is that the pseudo-true value θ_n needs to be indexed by sample size. (This idea is used in Hal White's work.)

Never-the-less, Schennach's theorem suggests that empirical likelihood is non-robust to misspecification.

15.17 Exponential Tilting

Instead of

$$KLIC(f,g) = \int f(y) \log \left(\frac{f(y)}{g(y)} \right) dy$$

consider the reverse distance

$$KLIC(g,f) = \int g(y) \log \left(\frac{g(y)}{f(y)} \right) dy.$$

The pseudo-true g which minimizes this criterion is

$$\min_g \int g(y) \log \left(\frac{g(y)}{f(y)} \right) dy$$

subject to

$$\begin{aligned} \int g(y)dy &= 1 \\ \int m(y)g(y)dy &= 0 \end{aligned}$$

The Lagrangian is

$$\int g(y) \log \left(\frac{g(y)}{f(y)} \right) dy - \mu \left(\int g(y)dy - 1 \right) - \lambda' \int m(y)g(y)dy$$

with FOC

$$0 = \log \left(\frac{g(y)}{f(y)} \right) + 1 - \mu - \lambda' m(y).$$

Solving

$$g(y) = f(y) \exp \left(-1 + \mu \right) \exp \left(\lambda' m(y) \right).$$

Imposing $\int g(y)dy = 1$ we find

$$g(y) = \frac{f(y) \exp \left(\lambda' m(y) \right)}{\int f(y) \exp \left(\lambda' m(y) \right) dy}. \tag{10}$$

Hence the name "exponential tilting" or ET

Inserting this into $KLIC(g,f)$ we find

$$\begin{aligned}
KLIC(g,f) &= \int g(y)\log\left(\frac{\exp\left(\lambda' m(y)\right)}{\int f(y)\exp\left(\lambda' m(y)\right)dy}\right)dy \\
&= \lambda'\int m(y)g(y)dy - \int g(y)dy\log\left(\int f(y)\exp\left(\lambda' m(y)\right)dy\right) \\
&= -\log\left(\int f(y)\exp\left(\lambda' m(y)\right)dy\right) \qquad (11) \\
&= -\log E\exp\left(\lambda' m(y)\right) \qquad (12)
\end{aligned}$$

By duality, the optimal Lagrange multiplier λ_0 maximizes this expression, equivalently

$$\lambda_0 = \underset{\lambda}{\operatorname{argmin}}\, E\exp\left(\lambda' m(y)\right) \qquad (13)$$

The pseudo-true density $g_0(y)$ is (10) with this λ_0, with associated minimized KLIC (11). This is the smallest possible KLIC(g,f) for moment condition models.

Notice: the g_0 which minimize KLIC(g,f) and KLIC(f,g) are different.

In contrast to the EL case, the ET problem (13) does not restrict λ, and there are no "trouble spots". Thus ET is more robust than EL. The pseudo-true λ_0 and g_0 are well defined under misspecification, unlike EL.

When the moment $m(y,\theta)$ depends on a parameter θ,then the pseudo-true values (θ_0,λ_0) are the joint solution to the problem

$$\max_{\theta}\min_{\lambda} E\exp\left(\lambda' m(y,\theta)\right).$$

15.18 Exponential Tilting – Estimation

The ET or exponential tilting estimator solves the problem

$$\min_{\theta,p_1,\ldots,p_n}\sum_{i=1}^{n} p_i\log p_i$$

subject to

$$\begin{aligned}
\sum_{i=1}^{n} p_i &= 1 \\
\sum_{i=1}^{n} p_i m\left(y_i,\theta\right) &= 0
\end{aligned}$$

First, we concentrate out the probabilities. For any θ, the Lagrangian is

$$\sum_{i=1}^{n} p_i \log p_i - \mu \left(\sum_{i=1}^{n} p_i - 1 \right) - \lambda' \sum_{i=1}^{n} p_i m \left(y_i, \theta \right)$$

with FOC

$$0 = \log \hat{p}_i - 1 - \mu - \lambda' m(y_i, \theta).$$

Solving for $\hat{p}_i$ and imposing the summability,

$$\hat{p}_i(\lambda) = \frac{\exp\left(\lambda' m(y_i, \theta)\right)}{\sum_{i=1}^{n} \exp\left(\lambda' m(y_i, \theta)\right)}$$

When $\lambda = 0$ then $\hat{p}_i = n^{-1}$, same as EL. The concentrated "entropy" criterion is then

$$\begin{aligned} \sum_{i=1}^{n} \hat{p}_i(\lambda) \log \hat{p}_i(\lambda) &= \sum_{i=1}^{n} \hat{p}_i(\lambda) \left[\lambda' m(y_i, \theta) - \log \left(\sum_{i=1}^{n} \exp\left(\lambda' m(y_i, \theta)\right) \right) \right] \\ &= -\log \left(\sum_{i=1}^{n} \exp\left(\lambda' m(y_i, \theta)\right) \right) \end{aligned}$$

By duality, the Lagrange multiplier maximizes this criterion, or equivalently

$$\hat{\lambda}(\theta) = \underset{\lambda}{\operatorname{argmin}} \sum_{i=1}^{n} \exp\left(\lambda' m(y_i, \theta)\right)$$

The ET estimator $\hat{\theta}$ maximizes this concentrated function, e.g.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \exp\left(\hat{\lambda}(\theta)' m(y_i, \theta)\right)$$

The ET probabilities are $\hat{p}_i = \hat{p}_i(\hat{\lambda})$

15.19 Schennach's Estimator

Schennach (2007) observed that while the ET probabilities have desirable properties, the EL estimator for θ has better bias properties. She suggested a hybrid estimator which achieves the best of both worlds, called exponentially tilted empirical likelihood (ETEL).

This is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, ELET(\theta)$$

$$\begin{aligned} ETEL(\theta) &= \sum_{i=1}^{n} \log\left(\hat{p}_i(\theta)\right) \\ &= \hat{\lambda}(\theta)' \sum_{i=1}^{n} m(y_i, \theta) - \log\left(\sum_{i=1}^{n} \exp\left(\hat{\lambda}(\theta)' m(y_i, \theta)\right)\right) \end{aligned}$$

$$\hat{p}_i(\theta) = \frac{\exp\left(\hat{\lambda}(\theta)' m(y_i, \theta)\right)}{\sum_{i=1}^{n} \exp\left(\hat{\lambda}(\theta)' m(y_i, \theta)\right)}$$

$$\hat{\lambda}(\theta) = \underset{\lambda}{\operatorname{argmin}} \sum_{i=1}^{n} \exp\left(\lambda' m(y_i, \theta)\right)$$

She claims the following advantages for the ETEL estimator $\hat{\theta}$

- Under correct specification, $\hat{\theta}$ is asymptotically second-order equivalent to EL
- Under misspecification, the pseudo-true values λ_0, θ_0 are generically well defined, and minimize a KLIC analog
- $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \rightarrow_d N\left(0, \Gamma^{-1}\Omega\Gamma^{-1\prime}\right)$ where $\Gamma = E\frac{\partial}{\partial\theta'} m\left(y, \theta\right)$ and $\Omega = E\ m\left(y, \theta\right) m\left(y, \theta\right)'$.

16 Model Averaging

16.1 Framework

Let g be a (non-parametric) object of interest, such as a conditional mean, variance, density, or distribution function. Let $\hat{g}_m$, $m = 1, ..., M$ be a discrete set of estimators. Most commonly, this set is the same as we might consider for the problem of model selection. In linear regression, typically $\hat{g}_m$ correspond to different sets of regressors. We will sometimes call the m'th estimator the m'th "model".

Let w_m be a set of weights for the m'th estimator. Let $\mathbf{w} = (w_1, ..., w_M)$ be the vector of weights. Typically we will require

$$\begin{aligned} 0 &\leq w_m \leq 1 \\ \sum_{m=1}^{M} w_m &= 1 \end{aligned}$$

The set of weights satisfying this condition is H_M, the unit simplex in $\mathbb{R}^M$.

An averaging estimator is

$$\hat{g}(\mathbf{w}) = \sum_{m=1}^{M} w_m \hat{g}_m$$

It is commonly called a "model average estimator".

Selection estimators are the special case where we impose the restriction $w_m \in \{0, 1\}$.

16.2 Model Weights

The most common method for weight specification is Bayesian Model Averaging (BMA). Assume that there are M potential models and one of the models is the true model. Specify prior probabilities that each of the potential models is the true model. For each model specify a prior over the parameters. Then the posterior distribution is the weighted average of the individual models, where the weights are Bayesian posterior probabilities that the given model is the true model, conditional on the data.

Given diffuse priors and equal model prior probabilities, the BMA weights are approximately

$$w_m = \frac{\exp\left(-\frac{1}{2} BIC_m\right)}{\sum_{j=1}^{M} \exp\left(-\frac{1}{2} BIC_j\right)}$$

where

$$BIC_m = 2\mathcal{L}_m + k_m \log(n)$$

$\mathcal{L}_m$ is the negative log-likelihood, and k_m is the number of parameters in model m. BIC_m is the Bayesian information criterion for model m. It is similar to AIC, but with the "2" replaced by log(n).

The BMA estimator has the nice interpretation as a Bayesian estimator. The downside is that it does not allow for misspecification. It is designed to search for the "true" model, not to select an estimator with low loss.

To remedy this situation, Burnham and Anderson have suggested replacing BIC with AIC, resulting in what has been called smoothed AIC (AIC) or weighted AIC (WAIC). The weights are

$$w_m = \frac{\exp\left(-\frac{1}{2}AIC_m\right)}{\sum_{j=1}^{M}\exp\left(-\frac{1}{2}AIC_j\right)}$$

where

$$AIC_m = 2\mathcal{L}_m + 2k_m$$

The suggestion goes back to Akaike, who suggested that these w_m may be interpreted as model probabilities. It is convenient and simple to implement. The idea can be applied quite broadly, in any context where AIC is defined.

In simulation studies, the SAIC estimator performs very well. (In particular, better than conventional AIC.) However, to date I have seen no formal justification for the procedure. It is unclear in what sense SAIC is producing a good approximation.

16.3 Linear Regression

In the case of linear regression, let X_m be regressor matrix for the m'th estimator. Then the list of all regressors. Then the m'th estimator is

$$\begin{aligned}\hat{\beta}_m &= \left(X_m'X_m\right)^{-1}X_my \\ \hat{g}_m &= X_m\hat{\beta}_m \\ &= P_my\end{aligned}$$

where

$$P_m = X_m\left(X_m'X_m\right)^{-1}X_m$$

The averaging estimator is

$$\begin{aligned}\hat{g}\left(\mathbf{w}\right) &= \sum_{m=1}^{M} w_m\hat{g}_m \\ &= \sum_{m=1}^{M} w_mP_my \\ &= P\left(\mathbf{w}\right)y\end{aligned}$$

where

$$P(\mathbf{w}) = \sum_{m=1}^{M} w_m P_m$$

Let X be the matrix of all regressors. We can also write

$$\begin{aligned}\hat{g}(\mathbf{w}) &= \sum_{m=1}^{M} w_m X_m \left(X_m' X_m\right)^{-1} X_m y \\ &= \sum_{m=1}^{M} w_m X_m \hat{\beta}_m \\ &= X \sum_{m=1}^{M} w_m \begin{pmatrix} \hat{\beta}_m \\ 0 \end{pmatrix} \\ &= X \hat{\beta}(\mathbf{w})\end{aligned}$$

where

$$\hat{\beta}(\mathbf{w}) = \sum_{m=1}^{M} w_m \begin{pmatrix} \hat{\beta}_m \\ 0 \end{pmatrix}$$

is the average of the coefficient estimates. $\hat{\beta}(\mathbf{w})$ is the model average estimator for β. In linear regression, there is a direct correspondence between the average estimator for the conditional mean and the average estimator of the parameters, but this correspondence breaks down when the estimator is not linear in the parameters.

16.4 Mallows Weight Selection

As pointed out above, in the linear regression setting, $\hat{g}(\mathbf{w}) = P(\mathbf{w})\, y$ is a linear estimator, so falls in the class studied by Li (1987). His framework allows for estimators indexed by $\mathbf{w} \in H_M$

Under homoskedasticity, an optimal method for selection of $\mathbf{w}$ is the Mallows criterion. As we discussed before, for estimators $\hat{g}(\mathbf{w}) = P(\mathbf{w})\, y$, the Mallows criterion is

$$C(\mathbf{w}) = \hat{e}(\mathbf{w})' \hat{e}(\mathbf{w}) + 2\sigma^2 \operatorname{tr} P(\mathbf{w})$$

where

$$\hat{e}(\mathbf{w}) = y - \hat{g}(\mathbf{w})$$

is the residual.

In averaging linear regression

$$\begin{aligned}
\operatorname{tr} P(\mathbf{w}) &= \operatorname{tr} \sum_{m=1}^{M} w_m P_m \\
&= \sum_{m=1}^{M} w_m \operatorname{tr} P_m \\
&= \sum_{m=1}^{M} w_m k_m \\
&= \mathbf{w}'\mathbf{K}
\end{aligned}$$

where k_m is the number of coefficients in the m'th model, and $\mathbf{K} = (k_1, ..., k_M)'$. The penalty is twice $\mathbf{w}'\mathbf{K}$, the (weighted) average number of coefficients.

Also

$$\begin{aligned}
\hat{e}(\mathbf{w}) &= y - \hat{g}(\mathbf{w}) \\
&= \sum_{m=1}^{M} w_m (y - \hat{g}_m) \\
&= \sum_{m=1}^{M} w_m \hat{e}_m \\
&= \hat{\mathbf{e}}\mathbf{w}
\end{aligned}$$

where $\hat{e}_m$ is the $n \times 1$ residual vector from the m'th model, and $\hat{\mathbf{e}} = [\hat{e}_1, ..., \hat{e}_M]$ is the $n \times M$ matrix of residuals from all M models.

We can then write the criterion as

$$C(\mathbf{w}) = \mathbf{w}'\hat{\mathbf{e}}'\hat{\mathbf{e}}\mathbf{w} + 2\sigma^2\mathbf{w}'\mathbf{K}$$

This is quadratic in the vector $\mathbf{w}$.

The Mallows selected weight vector minimizes the criterion $C(\mathbf{w})$ over $\mathbf{w} \in H_M$, the unit simplex.

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in H_M}{\operatorname{argmin}} C(\mathbf{w})$$

This is a quadratic programming problem with inequality constraints, which is pre-programmed in Gauss and Matlab, so computation of $\hat{\mathbf{w}}$ is a simple command.

The Mallows selected estimator is then

$$\begin{aligned}
\hat{g} &= \hat{g}(\hat{\mathbf{w}}) \\
&= \sum_{m=1}^{M} \hat{w}_m \hat{g}_m
\end{aligned}$$

This is an

16.5 Weight Selection Optimality

As we discussed in the section on model selection, Li (1987) provided a set of sufficient conditions for the Mallows selected estimator to be optimal, in the sense that the squared error is asymptotically equivalent to the infeasible optimum. The key condition was

$$\sum_{\mathbf{w}} (nR(\mathbf{w}))^{-s} \to 0 \tag{1}$$

In Hansen (Econometrica, 2007), I show that this condition is satisfied if we restrict the set of weights to a discrete set.

Recall that H_M is the unit simplex in $\mathbb{R}^M$.

Now restrict $\mathbf{w} \in H_M^* \subset H_M$, where the weights in H_M^* are elements of $\{\frac{1}{N}, \frac{2}{N}, ..., 1\}$ for some integer N. In that paper, I show that Li's condition (1) over $\mathbf{w} \in H_M^*$ holds under the similar conditions as model selection, namely if the models are nested,

$$\xi_n = \inf_{\mathbf{w} \in H_M} nR(\mathbf{w}) \to \infty$$

and

$$E\left(e_i^{4(N+1)} \mid X_i\right) \leq \kappa < \infty.$$

Thus model averaging is asymptotically optimal, in the sense that

$$\frac{L(\hat{\mathbf{w}})}{\inf_{\mathbf{w}\in H_M^*} L(\mathbf{w})} \to_p 1$$

where, again

$$L(\mathbf{w}) = \frac{1}{n}\left(\hat{\mathbf{g}}\left(\mathbf{w}\right) - \mathbf{g}\right)'\left(\hat{\mathbf{g}}\left(\mathbf{w}\right) - \mathbf{g}\right)$$

The proof is similar to that for model selection in linear regression. The restriction of $\mathbf{w}$ to a discrete set is necessary to directly apply Li's theorem, as the summation requires discreteness.

The discreteness was relaxed in a paper by Wan, Zhang, and Zou (2008, Least Squares Model Combining by Mallows Criterion, working paper). Rather than proving (1), they provided a more basic derivation, although using stronger conditions. Recall that the proof requires showing uniform convergence results of the form

$$\sup_{\mathbf{w}\in H_M} \frac{|e'b(\mathbf{w})|}{nR(\mathbf{w})} \to_p 0$$

where

$$\begin{aligned} b(\mathbf{w}) &= \sum_{m=1}^{M} w_m b_m \\ b_m &= (I - P_m)\,\mathbf{g} \end{aligned}$$

Here is their proof: First,

$$\sup_{\mathbf{w}\in H_M} \frac{|e'b(\mathbf{w})|}{nR(\mathbf{w})} \leq \sum_{m=1}^{M} w_m \frac{|e'b_m|}{\xi_n} \leq \max_{1\leq m\leq M} \frac{|e'b_m|}{\xi_n}$$

Second, by Markov's and Whittle's inequalities

$$\begin{aligned} P\left(\max_{1\leq m\leq M} \frac{|e'b_m|}{\xi_n} > \delta\right) &\leq \sum_{m=1}^{M} P\left(\frac{|e'b_m|}{\xi_n} > \delta\right) \\ &\leq \sum_{m=1}^{M} \frac{E\,|e'b_m|^{2G}}{\delta^{2G}\xi_n^{2G}} \\ &\leq K\sum_{m=1}^{M} \frac{|b_m'b_m|^{G}}{\delta^{2G}\xi_n^{2G}} \\ &\leq K\sum_{m=1}^{M} \frac{\left(nR(\mathbf{w}_m^0)\right)^{G}}{\delta^{2G}\xi_n^{2G}} \end{aligned}$$

where $\mathbf{w}_m^0$ is the weight vector with a 1 in the m'th place and zeros elsewhere. Equivalently, $nR(\mathbf{w}_m^0)$ is the expected squared error from the m'th model. The final inequality uses the fact from the analysis for model selection that

$$nR(\mathbf{w}_m^0) = b_m'b_m + \sigma^2 k_m \leq b_m'b_m$$

Wan, Zhang, and Zou then assume

$$\frac{\sum_{m=1}^{M} \left(nR(\mathbf{w}_m^0)\right)^{G}}{\xi_n^{2G}} \to 0$$

This is stronger than the condition from my paper $\xi_n \to \infty$, as it requires that $\sum_{m=1}^{M} \left(nR(\mathbf{w}_m^0)\right)^{G}$ diverges slower than ξ_n^{2G}. They also do not directly assume that the models are nested.

16.6 Cross-Validation Selection

Hansen and Racine (Jacknife Model Averaging, working paper).

In this paper, we substitute CV for the Mallows criterion. As a result, we do not require homoskedasticity.

For the $m'th$ model, let $\tilde{e}_i^m$ denote the leave-one-out (LOO) residuals for the i'th observation, e.g.

$$\tilde{e}_i^m = y_i - X_i^{m\prime} \left(\mathbf{X}_{-i}^{m\prime} \mathbf{X}_{-i}^m \right)^{-1} \mathbf{X}_{-i}^{m\prime} \mathbf{y}_{-i}$$

and let $\tilde{e}_m$ denote the $n \times 1$ vector of the $\tilde{e}_i^m$. Then the LOO averaging residuals are

$$\tilde{e}_i(\mathbf{w}) = \sum_{m=1}^M w_m \tilde{e}_i^m$$

$$\tilde{e}(\mathbf{w}) = \sum_{m=1}^M w_m \tilde{e}^m = \tilde{\mathbf{e}} \mathbf{w}$$

where $\tilde{\mathbf{e}}$ is an $n \times M$ matrix whose mth column is $\tilde{e}_m$. Then the sum-of-squared LOO residuals is

$$CV(\mathbf{w}) = \tilde{e}(\mathbf{w})' \tilde{e}(\mathbf{w}) = \mathbf{w}' \tilde{e}' \tilde{e} \mathbf{w}$$

which is quadratic in $\mathbf{w}$.

The CV (or jacknife) selected weight vector $\hat{\mathbf{w}}$ minimizes the criterion $CV(\mathbf{w})$ over the unit simplex. As for Mallows selection, this is solved by quadratic programming.

The JMA estimator is then $\hat{g}(\hat{\mathbf{w}})$

In Hansen-Racine, we show that the CV estimator is asymptotically equivalent to the infeasible best weight vector, under the conditions

$$\begin{aligned} 0 &< \min_i E\left(e_i^2 \mid X_i\right) \leq \min_i E\left(e_i^2 \mid X_i\right) < \infty \\ E\left(e_i^{4(N+1)} \mid X_i\right) &\leq \kappa < \infty \\ \xi_n &= \inf_{\mathbf{w} \in H_M} nR(\mathbf{w}) \to \infty \\ \max_{1 \leq m \leq M} \max_{1 \leq i \leq n} X_i^{m\prime} \left(\mathbf{X}^{m\prime} \mathbf{X}^m \right)^{-1} \mathbf{X}^{m\prime} X_i^m &\to 0 \end{aligned}$$

16.7 Many Unsolved Issues

- Model averaging for other estimators: e.g. densities or conditional densities
- IV, GMM, EL, ET
- Standard errors?
- Inference

17 Shrinkage

17.1 Mallows Averaging and Shrinkage

Suppose there are two models or estimators of $g = E(y \mid X)$

(1) $g_0 = 0$

(2) $\hat{g}_1 = X\hat{\beta}$

Given weights $(1-w)$ and w an averaging estimator is $\hat{g} = wX\hat{\beta}$.

The Mallows criterion is

$$\begin{aligned} C(w) &= \mathbf{w}'\hat{\mathbf{e}}'\hat{\mathbf{e}}\mathbf{w} + 2\hat{\sigma}^2\mathbf{w}'\mathbf{K} \\ &= \begin{pmatrix} 1-w & w \end{pmatrix} \begin{pmatrix} y'y & y'\hat{e} \\ \hat{e}'y & \hat{e}'\hat{e} \end{pmatrix} \begin{pmatrix} 1-w \\ w \end{pmatrix} + 2\hat{\sigma}^2 wk \\ &= (1-w)^2 y'y + \left(w^2 + 2w(1-w)\right)\hat{e}'\hat{e} + 2\hat{\sigma}^2 wk \\ &= (1-w)^2\left(y'y - \hat{e}'\hat{e}\right) + \hat{e}'\hat{e} + 2\hat{\sigma}^2 wk \end{aligned}$$

The FOC for minimization is

$$\frac{d}{dw}C(w) = -2(1-w)\left(y'y - \hat{e}'\hat{e}\right) + 2\hat{\sigma}^2 k = 0$$

with solution

$$\hat{w} = 1 - \frac{k}{F}$$

where

$$F = \frac{y'y - \hat{e}'\hat{e}}{\hat{\sigma}^2}$$

is the Wald statistic for $\beta = 0$. Imposing the constraint $\hat{w} \in [0,1]$ we obtain

$$\hat{w} = \begin{cases} 1 - \dfrac{k}{F} & F \geq k \\ \\ 0 & F < k \end{cases}$$

The Mallow averaging estimator thus equals

$$\hat{\beta}^* = \hat{\beta}\left(1 - \frac{k}{F}\right)_+$$

where $(a)_+ = a$ if $a \geq 0$, 0 else.

This is a Stein-type shrinkage estimator.

17.2 Loss and Risk

A great reference is *Theory of Point Estimation*, 2nd Edition, by Lehmann and Casella.

Let $\hat{\theta}$ be an estimator for θ, $k \times 1$. Suppose $\hat{\theta}$ is an (asymptotic) sufficient statistic for θ so that any other estimator can be written as a function of $\hat{\theta}$. We call $\hat{\theta}$ the "usual"estimator.

Suppose that $\sqrt{n}\left(\hat{\theta} - \theta\right) \to_d N(0, \mathbf{V})$. Thus, approximately,

$$\hat{\theta} \sim_a N(\theta, \mathbf{V}_n)$$

where $\mathbf{V}_n = n^{-1}\mathbf{V}$. Most of Stein-type theory is developed for the exact distribution case. It carries over to the asymptotic setting as approximations. For now on we will assume that $\hat{\theta}$ has an exact normal distribution, and that $\mathbf{V}_n = \mathbf{V}$ is known. (Equivalently, we can rewrite the statistical problem as local to θ using the "Limits of Experiments" theory.

Is $\hat{\theta}$ the best estimator for θ, in the sense of minimizing the risk (expected loss)?

The risk of $\tilde{\theta}$ under weighted squared error loss is

$$\begin{aligned} R(\theta, \tilde{\theta}, \mathbf{W}) &= E\left(\left(\tilde{\theta} - \theta\right)' \mathbf{W} \left(\tilde{\theta} - \theta\right)\right) \\ &= \operatorname{tr}\left(\mathbf{W} E\left(\left(\tilde{\theta} - \theta\right)\left(\tilde{\theta} - \theta\right)\right)'\right) \end{aligned}$$

A convenient choice for the weight matrix is $\mathbf{W} = \mathbf{V}^{-1}$. Then

$$\begin{aligned} R(\theta, \hat{\theta}, \mathbf{V}^{-1}) &= \operatorname{tr}\left(\mathbf{V}^{-1} E\left(\left(\tilde{\theta} - \theta\right)\left(\tilde{\theta} - \theta\right)\right)'\right) \\ &= \operatorname{tr}\left(\mathbf{V}^{-1}\mathbf{V}\right) \\ &= k. \end{aligned}$$

If $\mathbf{W} \neq \mathbf{V}^{-1}$ then

$$\begin{aligned} R(\theta, \hat{\theta}, \mathbf{W}) &= \operatorname{tr}\left(\mathbf{W} E\left(\left(\tilde{\theta} - \theta\right)\left(\tilde{\theta} - \theta\right)\right)'\right) \\ &= \operatorname{tr}\left(\mathbf{W}\mathbf{V}\right) \end{aligned}$$

which depends on $\mathbf{W}\mathbf{V}$.

Again, we want to know if the risk of another feasible estimator is smaller than $\operatorname{tr}\left(\mathbf{W}\mathbf{V}\right)$.

Take the simple (or silly) estimator $\tilde{\theta} = 0$. This has risk

$$R(\theta, 0, \mathbf{W}) = \theta' \mathbf{W} \theta.$$

Thus $\tilde{\theta} = 0$ has smaller risk than $\hat{\theta}$ when $\theta'\mathbf{W}\theta < \operatorname{tr}\left(\mathbf{W}\mathbf{V}\right)$, and larger risk when $\theta'\mathbf{W}\theta > \operatorname{tr}\left(\mathbf{W}\mathbf{V}\right)$. Neither $\hat{\theta}$ nor $\tilde{\theta} = 0$ is "better" in the sense of having (uniformly) smaller risk! It is not enough to ask that one estimator has smaller risk than another, as in general the risk is a function depending on unknowns.

As another example, take the simple averaging (or shrinkage) estimator

$$\tilde{\theta} = w\hat{\theta}$$

where w is a fixed constant. Since

$$\tilde{\theta} - \theta = w\left(\hat{\theta} - \theta\right) - (1-w)\theta$$

we can calculate that

$$\begin{aligned} R(\theta, \tilde{\theta}, \mathbf{W}) &= w^2 R(\theta, \tilde{\theta}, \mathbf{W}) + (1-w)^2\theta'\mathbf{W}\theta \\ &= w^2 \operatorname{tr}(\mathbf{WV}) + (1-w)^2\theta'\mathbf{W}\theta \end{aligned}$$

This is minimized by setting

$$w = \frac{\theta'\mathbf{W}\theta}{\operatorname{tr}(\mathbf{WV}) + \theta'\mathbf{W}\theta}$$

which is strictly in (0,1). [This is illustrative, and does not suggest an empirical rule for selecting w.]

17.3 Admissibile and Minimax Estimators

For reference.

To compare the risk functions of two estimators, we have the following concepts.

Definition 1 $\hat{\theta}$ *weakly dominates* $\tilde{\theta}$ *if* $R(\theta, \hat{\theta}) \leq R(\theta, \tilde{\theta})$ *for all* θ

Definition 2 $\hat{\theta}$ *dominates* $\tilde{\theta}$ *if* $R(\theta, \hat{\theta}) \leq R(\theta, \tilde{\theta})$ *for all* θ, *and* $R(\theta, \hat{\theta}) < R(\theta, \tilde{\theta})$ *for at least one* θ.

Clearly, we should prefer an estimator if it dominates the other.

Definition 3 *An estimator is admissible if it is not dominated by another estimator. An estimator is inadmissible if it is dominated by another estimator.*

Admissibility is a desirable property for an estimator.

If the risk functions of two estimators cross, then neither dominates the other. How do we compare these two estimators?

One approach is to calculate the worst-case scenerio. Specifically, we define the maximum risk of an estimator $\tilde{\theta}$ as

$$\bar{R}(\tilde{\theta}) = \sup_{\theta} R\left(\theta, \tilde{\theta}\right)$$

We can think: Suppose we use $\tilde{\theta}$ to estimate θ. Then what is the worst case, how bad can than this estimator do?

For example, for the usual estimator, $R\left(\theta, \hat{\theta}, \mathbf{W}\right) = \operatorname{tr}(\mathbf{WV})$ for all θ, so

$$\bar{R}(\hat{\theta}) = \operatorname{tr}(\mathbf{WV})$$

while for the silly estimator $\tilde{\theta} = 0$

$$\bar{R}(0) = \infty$$

The latter is an example of an estimator with unbounded risk. To guard against extreme worst cases, it seems sensible to avoid estimators with unbounded risk.

The minimium value of the maximum risk $\bar{R}(\tilde{\theta})$ across all estimators $\delta = \delta(\hat{\theta})$ is

$$\inf_{\delta} \bar{R}(\delta) = \inf_{\delta} \sup_{\theta} R(\theta, \delta)$$

where

Definition 4 *An estimator $\tilde{\theta}$ of θ which minimizes the maximum risk*

$$\inf_{\delta} \sup_{\theta} R(\theta, \delta) = \sup_{\theta} R(\theta, \tilde{\theta})$$

is called a minimax estimator.

It is desireable for an estimator to be minimax, again as a protection against the worst-case scenerio.

There is no general rule for determining the minimax bound. However, in the case $\hat{\theta} \sim N(\theta, I_k)$, it is known that $\hat{\theta}$ is mimimax for θ.

17.4 Shrinkage Estimators

Suppose $\hat{\theta} \sim N(\theta, \mathbf{V})$

A general form for a shrinkage estimator for θ is

$$\hat{\theta}^* = \left(1 - h(\hat{\theta}' \mathbf{W} \hat{\theta})\right) \hat{\theta}$$

where $h : [0, \infty) \to [0, \infty)$. Sometimes this is written as

$$\hat{\theta}^* = \left(1 - \frac{c(\hat{\theta}' \mathbf{W} \hat{\theta})}{\hat{\theta}' \mathbf{W} \hat{\theta}}\right) \hat{\theta}$$

where $c(q) = qh(q)$.

This notation includes the James-Stein estimator, pretest estimators, selection estimators, and the Model averaging estimator of section 17.1. Pretest and selection estimators take the form

$$h(q) = 1(q < a)$$

where $a = 2k$ for Mallows selection, and a is the critical value from a chi-square distribution for a pretest estimator.

We now calculate the risk of $\hat{\theta}^*$. Note

$$\hat{\theta}^* - \theta = \left(\hat{\theta} - \theta\right) - h(\hat{\theta}'\mathbf{W}\hat{\theta})\hat{\theta}$$

Thus

$$\left(\hat{\theta}^* - \theta\right)'\mathbf{W}\left(\hat{\theta}^* - \theta\right) = \left(\hat{\theta} - \theta\right)'\mathbf{W}\left(\hat{\theta} - \theta\right) + h(\hat{\theta}'\mathbf{W}\hat{\theta})^2\hat{\theta}'\mathbf{W}\hat{\theta} - 2h(\hat{\theta}'\mathbf{W}\hat{\theta})\hat{\theta}'\mathbf{W}\left(\hat{\theta} - \theta\right)$$

Taking expectations:

$$R(\theta, \hat{\theta}^*, \mathbf{W}) = \operatorname{tr}(\mathbf{W}\mathbf{V}) + E\left[h(\hat{\theta}'\mathbf{W}\hat{\theta})^2\hat{\theta}'\mathbf{W}\hat{\theta}\right] - 2E\left[h(\hat{\theta}'\mathbf{W}\hat{\theta})\hat{\theta}'\mathbf{W}\left(\hat{\theta} - \theta\right)\right]$$

To simplify the second expectation when h is continuous we use:

Lemma 1 *(Stein's Lemma) If $\eta(\theta) : \mathbb{R}^k \to \mathbb{R}^k$ is absolutely continuous and $\hat{\theta} \sim N(\theta, \mathbf{V})$ then*

$$\mathrm{E}\left(\eta(\hat{\theta})'\left(\hat{\theta} - \theta\right)\right) = \mathrm{E}\operatorname{tr}\left(\frac{\partial}{\partial\theta}\eta(\hat{\theta})'\mathbf{V}\right).$$

Proof: Let

$$\phi(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}}\exp\left(-\frac{1}{2}\mathbf{x}'\mathbf{V}^{-1}\mathbf{x}\right)$$

denote the $\mathrm{N}(\mathbf{0}, \mathbf{V})$ density. Then

$$\frac{\partial}{\partial\mathbf{x}}\phi(\mathbf{x}) = -\mathbf{V}^{-1}\mathbf{x}\phi(\mathbf{x})$$

and

$$\frac{\partial}{\partial\mathbf{x}}\phi(\mathbf{x} - \theta) = -\mathbf{V}^{-1}(\mathbf{x} - \theta)\phi(\mathbf{x} - \theta).$$

By multivariate integration by parts

$$\begin{aligned}
\mathrm{E}\left(\eta(\hat{\theta})'\left(\hat{\theta} - \theta\right)\right) &= \int \eta(\mathbf{x})'\mathbf{V}\mathbf{V}^{-1}(\mathbf{x} - \theta)\,\phi(\mathbf{x} - \theta)\,(\mathbf{dx}) \\
&= \int \operatorname{tr}\left(\frac{\partial}{\partial\theta}\eta(\mathbf{x})'\mathbf{V}\phi(\mathbf{x} - \theta)\right)(\mathbf{dx}) \\
&= \mathrm{E}\operatorname{tr}\left(\frac{\partial}{\partial\theta}\eta(\hat{\theta})'\mathbf{V}\right)
\end{aligned}$$

as stated. ■

Let $\eta(\theta)' = h\left(\theta'\mathbf{W}\theta\right)\theta'\mathbf{W}$, for which

$$\frac{\partial}{\partial\theta}\eta(\theta)' = h(\theta'\mathbf{W}\theta)\mathbf{W} + 2\mathbf{W}\theta\theta'\mathbf{W}h'(\theta'\mathbf{W}\theta)$$

and

$$\operatorname{tr} \frac{\partial}{\partial \theta} \eta(\theta)' \mathbf{V} = \operatorname{tr}(\mathbf{WV})\, h(\theta' \mathbf{W} \theta) + 2\theta' \mathbf{WVW} \theta h'(\theta' \mathbf{W} \theta)$$

Then by Stein's Lemma

$$E\left[h(\hat{\theta}' \mathbf{W} \hat{\theta}) \hat{\theta}' \mathbf{W} \left(\hat{\theta} - \theta\right)\right] = \operatorname{tr}(\mathbf{WV})\, Eh(\hat{\theta}' \mathbf{W} \hat{\theta}) + 2E\left[(\hat{\theta}' \mathbf{WVW} \hat{\theta}) h'(\hat{\theta}' \mathbf{W} \hat{\theta})\right]$$

Applying this to the risk calculation, we obtain

Theorem.

$$\begin{aligned} R(\theta, \hat{\theta}^*, \mathbf{W}) &= \operatorname{tr}(\mathbf{WV}) + E\left[h(\hat{\theta}' \mathbf{W} \hat{\theta})^2 \hat{\theta}' \mathbf{W} \hat{\theta}\right] - 2 \operatorname{tr}(\mathbf{WV})\, Eh(\hat{\theta}' \mathbf{W} \hat{\theta}) - 4E\left[(\hat{\theta}' \mathbf{WVW} \hat{\theta}) h'(\hat{\theta}' \mathbf{W} \hat{\theta})\right] \\ &= \operatorname{tr}(\mathbf{WV}) + E\left[c(\hat{\theta}' \mathbf{W} \hat{\theta}) \frac{\left(c(\hat{\theta}' \mathbf{W} \hat{\theta}) - 2 \operatorname{tr}(\mathbf{WV}) + 4 \frac{\hat{\theta}' \mathbf{WVW} \hat{\theta}}{\hat{\theta}' \mathbf{W} \hat{\theta}}\right)}{\hat{\theta}' \mathbf{W} \hat{\theta}} - 4 \frac{\hat{\theta}' \mathbf{WVW} \hat{\theta}}{\hat{\theta}' \mathbf{W} \hat{\theta}} c'(\hat{\theta}' \mathbf{W} \hat{\theta})\right] \end{aligned}$$

where the final equality uses the alternative expression $h(q) = c(q)/q$.

We are trying to find cases where $R(\theta, \hat{\theta}^*, \mathbf{W}) < R(\theta, \hat{\theta}, \mathbf{W})$. This requires the term in the expectation to be negative.

We now explore some special cases.

17.5 Default Weight Matrix

Set

$$\mathbf{W} = \mathbf{V}^{-1}$$

and write

$$R(\theta, \hat{\theta}^*) = R(\theta, \hat{\theta}^*, \mathbf{V}^{-1})$$

Then

$$R(\theta, \hat{\theta}^*) = k + E\left[c(\hat{\theta}' \mathbf{V}^{-1} \hat{\theta}) \frac{\left(c(\hat{\theta}' \mathbf{V}^{-1} \hat{\theta}) - 2k + 4\right)}{\hat{\theta}' \mathbf{V}^{-1} \hat{\theta}} - 4c'(\hat{\theta}' \mathbf{V}^{-1} \hat{\theta})\right].$$

Theorem 1 *For any absolutely continuous and non-decreasing function $c(q)$ such that*

$$0 < c(q) < 2(k-2) \tag{1}$$

then

$$R(\theta, \hat{\theta}^*) < R(\theta, \hat{\theta}),$$

the risk of $\hat{\theta}^*$ is strictly less than the risk of $\hat{\theta}$. This inequality holds for all values of the parameter θ.

Note: Condition (1) can only hold if $k > 2$. (Since k , the dimension of θ, is an integer, this means $k \geq 3$.)

Proof. Let

$$g(q) = \frac{c(q)\left(c(q) - 2\left(k-2\right)\right)}{q} - 4c'(q)$$

For all $q \geq 0$, $g(q) < 0$ by the assumptions. Thus $Eg\left(q\right)$ for any non-negative random variable q. Setting $q_k = \hat{\theta}'\mathbf{V}^{-1}\hat{\theta}$,

$$R(\theta, \hat{\theta}^*) = k + Eg\left(q_k\right) < k = R(\theta, \hat{\theta})$$

which proves the result.

It also useful to note that

$$q_k = \hat{\theta}'\mathbf{V}^{-1}\hat{\theta} \sim \chi^2_k\left(\psi\right)$$

a non-central chi-square random variable with k degrees of freedom and non-centrality parameter

$$\psi = \theta'\mathbf{V}^{-1}\theta$$

17.6 James-Stein Estimator

Set $c(q) = c$, a constant. This is the James-Stein estimator

$$\hat{\theta}^* = \left(1 - \frac{c}{\hat{\theta}'\mathbf{V}^{-1}\hat{\theta}}\right)\hat{\theta} \tag{2}$$

Theorem 2 *If* $\hat{\theta} \sim N(\theta, \mathbf{V})$, $k > 2$, *and* $0 < c < 2(k-2)$, *then for (2),*

$$R(\theta, \hat{\theta}^*) < R(\theta, \hat{\theta})$$

the risk of the James-Stein estimator is strictly less than the usual estimator. This inequality holds for all values of the parameter θ.

Since the risk is quadratic in c, we can also see that the risk is minimized by setting $c = k-2$. This yields the classic form of the James-Stein estimator

$$\hat{\theta}^* = \left(1 - \frac{k-2}{\hat{\theta}'\mathbf{V}^{-1}\hat{\theta}}\right)\hat{\theta}$$

17.7 Positive-Part James-Stein

If $\hat{\theta}'\mathbf{V}^{-1}\hat{\theta} < c$ then

$$1 - \frac{c}{\hat{\theta}'\mathbf{V}^{-1}\hat{\theta}} < 0$$

and the James-Stein estimator over-shrinks, and flips the sign of $\hat{\theta}^*$ relative to $\hat{\theta}$. This is corrected by using the positive-part version

$$\begin{aligned} \hat{\theta}^+ &= \left(1 - \frac{c}{\hat{\theta}'\mathbf{V}^{-1}\hat{\theta}}\right)_+ \hat{\theta} \\ &= \begin{cases} \left(1 - \frac{c}{\hat{\theta}'\mathbf{V}^{-1}\hat{\theta}}\right)\hat{\theta} & \hat{\theta}'\mathbf{V}^{-1}\hat{\theta} \geq c \\ 0 & \text{else} \end{cases} \end{aligned}$$

This bears some resemblance to selection estimators.

The positive-part estimator takes the shrinkage form with

$$c(q) = \begin{cases} c & q \geq c \\ q & q < c \end{cases}$$

or

$$h(q) = \begin{cases} \frac{c}{q} & q \geq c \\ 1 & q < c \end{cases}$$

In general the positive-part version of

$$\hat{\theta}^* = \left(1 - h(\hat{\theta}'\mathbf{W}\hat{\theta})\right)\hat{\theta}$$

is

$$\hat{\theta}^+ = \left(1 - h(\hat{\theta}'\mathbf{W}\hat{\theta})\right)_+ \hat{\theta}$$

Theorem. For any shrinakge estimator, $R(\theta, \hat{\theta}^+) < R(\theta, \hat{\theta})$

The proof is a bit technical, so we will skip it.

17.8 General Weight Matrix

Recall that for general $c(q)$ and weight $\mathbf{W}$ we had

$$R(\theta, \hat{\theta}^*, \mathbf{W}) = \operatorname{tr}(\mathbf{WV}) + E\left[c(\hat{\theta}'\mathbf{W}\hat{\theta})\frac{\left(c(\hat{\theta}'\mathbf{W}\hat{\theta}) - 2\operatorname{tr}(\mathbf{WV}) + 4\frac{\hat{\theta}'\mathbf{WVW}\hat{\theta}}{\hat{\theta}'\mathbf{W}\hat{\theta}}\right)}{\hat{\theta}'\mathbf{W}\hat{\theta}} - 4\frac{\hat{\theta}'\mathbf{WVW}\hat{\theta}}{\hat{\theta}'\mathbf{W}\hat{\theta}}c'(\hat{\theta}'\mathbf{W}\hat{\theta})\right]$$

Using a result about eigenvalues and setting $h = \mathbf{W}^{-1/2}\theta$

$$\begin{aligned}\frac{\hat{\theta}'\mathbf{WVW}\hat{\theta}}{\hat{\theta}'\mathbf{W}\hat{\theta}} &\leq \max_{\theta} \frac{\theta'\mathbf{WVW}\theta}{\theta'\mathbf{W}\theta} \\ &= \max_{h} \frac{h'\mathbf{W}^{1/2}\mathbf{V}\mathbf{W}^{1/2}h}{h'h} \\ &= \lambda_{\max}(\mathbf{W}^{1/2}\mathbf{V}\mathbf{W}^{1/2}) \\ &= \lambda_{\max}(\mathbf{WV})\end{aligned}$$

Thus if $c'(q) \geq 0$,

$$\begin{aligned}R(\theta, \hat{\theta}^*, \mathbf{W}) &\leq \operatorname{tr}(\mathbf{WV}) + E\left[c(\hat{\theta}'\mathbf{W}\hat{\theta})\frac{\left(c(\hat{\theta}'\mathbf{W}\hat{\theta}) - 2\operatorname{tr}(\mathbf{WV}) + 4\lambda_{\max}(\mathbf{WV})\right)}{\hat{\theta}'\mathbf{W}\hat{\theta}}\right] \\ &< \operatorname{tr}(\mathbf{WV})\end{aligned}$$

the final inequality if

$$0 < c(q) < 2\left(\operatorname{tr}(\mathbf{WV}) - 2\lambda_{\max}(\mathbf{WV})\right) \qquad (3)$$

When $W = V$, the upper bound is $2(k-2)$ so this is the same as for the default weight matrix.

Theorem 3 *For any absolutely continuous and non-decreasing function $c(q)$ such that (3) holds, then*

$$R(\theta, \hat{\theta}^*, \mathbf{W}) < R(\theta, \hat{\theta}, \mathbf{W}),$$

the risk of $\hat{\theta}^*$ is strictly less than the risk of $\hat{\theta}$.

17.9 Shrinkage Towards Restrictions

The classic James-Stein estimator shrinks towards the zero vector. More generally, shrinkage can be towards restricted estimators, or towards linear or non-linear subspaces.

These estimators take the form

$$\hat{\theta}^* = \hat{\theta} - h(\left(\hat{\theta} - \tilde{\theta}\right)'\mathbf{W}\left(\hat{\theta} - \tilde{\theta}\right))\left(\hat{\theta} - \tilde{\theta}\right)$$

where $\hat{\theta}$ is the unrestricted estimator (e.g. the long regression) and $\tilde{\theta}$ is the restricted estimator (e.g. the short regression).

The classic form is

$$\hat{\theta}^* = \hat{\theta} - \left(\frac{r-2}{\left(\hat{\theta} - \tilde{\theta}\right)'\hat{\mathbf{V}}^{-1}\left(\hat{\theta} - \tilde{\theta}\right)}\right)_1 \left(\hat{\theta} - \tilde{\theta}\right)$$

where $(a)_1 = \max(a, 1)$, $\hat{\mathbf{V}}$ is the covariance matrix for $\hat{\theta}$, and r is the number of restrictions (by the restriction from $\hat{\theta}$ to $\tilde{\theta}$).

This estimator shrinks $\hat{\theta}$ towards $\tilde{\theta}$, with the degree of shrinkage depending on the magnitude of $\left(\hat{\theta} - \tilde{\theta}\right)$.

This approach works for nested models, so that $\left(\hat{\theta} - \tilde{\theta}\right)' \hat{\mathbf{V}}^{-1} \left(\hat{\theta} - \tilde{\theta}\right)$ is approximately (noncentral) chi-square.

It is unclear how to extend the idea to non-nested models, where $\left(\hat{\theta} - \tilde{\theta}\right)' \hat{\mathbf{V}}^{-1} \left(\hat{\theta} - \tilde{\theta}\right)$ is not chi-square.

17.10 Inference

We discussed shrinkage estimation.

Model averaging, Selection, and Shrinkage estimators have non-standard non-normal distributions.

Standard errors, testing, and confidence intervals need development.