Nonparametric Estimation of Conditional Covariance Matrices

By

Soonmook (Simon) Lee

Department of Mathematics and Statistics

Portland State University, Portland, Oregon 97201, USA

Spring, 2021

In partial fulfillment of the requirements

for the Master of Science in Statistics

Advisor: Professor Ge Zhao

Second Reader: Professor Subhash Kochar

1 Introduction

Covariance matrices are basic resources that are used to estimate many linear models such as regression analysis, factor analysis, and structural equation modeling. In a covariance matrix we observe variances and covariances. Often we are concerned with heteroscedasticity or local variability of data, which can be described using conditional variances. Heteroscedasticity is a violation of the usual assumption in regression analyses. In this case, the variance function must be estimated to understand the behavior of variables of interest (eg., Andersen and Lund, 1997; Holst, Hössjer, Björklund, Ragnarson, and Edner, 1996; Yao and Tong, 1994). In other examples, we have interest in the variance function[1] itself in its own right (Box, 1988; Ruppert, Wand, Holst, and Hössjer, 1997). In similar context we are interested in studies of conditional covariances. Examples are in the literature of machine learning (Bilmes, 2000), risk management (Ledoit and Wolf, 2004), graphical modeling (Drton and Perlman, 2004; Edwards, 2000), and longitudinal data analysis (Diggle and Verbyla, 1998; Smith and Kohn, 2002).

When the covariance matrices vary along one or more covariates, it is required to investigate the nature of conditionality and analyze conditional covariance matrices instead of unconditional covariance matrices. This issue of conditionality in estimating covariance matrices accompanies any linear structural modeling as the covariances represent linear association among variables.

In most linear models regression models are a building block to construct more complicated models. In this research regression methods and conditional covariance matrices are approached from a nonparametric perspective as an extension of studies using nonparametric regression functions to estimate conditional means and variances. As many nonparametric regression estimators are linear smoothers, combining observed reponses linearly, it is natural that density estimation methods are building blocks of nonparametric regression methods. The relationship between density estimator and regression estima-

[1]Let $(X_i, Y_i), \cdots, (X_n, Y_n)$ be a sample of random pair and assume that they satisfy the heteroscedastistic nonparametric regression model: $y_i = m(x_i) + \varepsilon_i$, for $1 \leq i \leq n$, where $\varepsilon_i \sim f(0, v(x_i))$.The m is the mean function and v is the variance function.

tor $\hat{m}(x)$ can be expressed as $\hat{m}(x) = \sum_{i=1}^{n} y_i g_i(x)$, where y_i's are response data and $g_i(x)$ represents the weight for a fixed point x of a random variable X. The weight is actually a choice of kernel density.

In Section 2 of this paper, development of density estimator and its theoretical properties will be discussed, including its relationship to the regression estimator. In Section 3, conditional covariance estimator will be discussed as an extension of nonparametric regression analysis. In Section 4, a simulation study is conducted to see how the covariance estimator works. In Section 5, the covariance estimator will be applied to real data, the Boston Housing Data Set (Harrison & Rubinfeld, 1978). An overview of Yin et al's limitations and the need for further studies are detailed in Section 6.

2 Development of Kernel Estimator

2.1 From Histograms to Averaged Shifted Histograms

As the histogram provides visual information of a density function through frequencies or relative frequencies, use of histograms has gone through many stages of development from frequency histogram (limit: lack of normalization) to density histogram (limit: discontinuity in the histogram) and frequency polygons (limit: effect of bin origin is a nuisance parameter). Finally, a simple device has been created for resolving the bin edge problem of the frequency polygon with the computational merit retained in estimating a density based on bin counts. Faced with the challenge of choosing among the collection of multivariate frequency polygons, a solution would be to average several of the shifted frequency polygons (Scott, 1983, 1985). The resulting curve appears to be a frequency polygon as well in the same way as the average of piecewise linear curves is also piecewise linear. The resulting averaged shifted frequency polygon can be made nonnegative and integrating to 1 by adjusting the weights.

However, it is simpler to average several shifted histograms [2] retaining the same generality as the averaged shifted frequency polygon. The "Averaged Shifed Histogram" (ASH) is more practical with computational and statistical efficiency in density estimation (Scott, 2015). Considering a collection of s histograms, $\hat{f}_1, \hat{f}_2, \cdots, \hat{f}_s$, each with the same bin width h_b, but with different bin origins $a_0 = 0, \frac{h_b}{s}, \frac{2h_b}{s}, \cdots, \frac{(s-1)h_b}{s}$, the unweighted ASH is written as

$$\hat{f}(\cdot) = \hat{f}_{ASH}(\cdot) = \frac{1}{s}\sum_{i=1}^{s} \hat{f}_i(\cdot), \tag{1}$$

where s is the number of histograms.

Univariate ASH is piecewise constant over the narrower intervals $[k\delta, (k\delta + \delta))$ with $\delta \equiv \frac{h_b}{s}$. And bin origins differ by the size of δ. Then the k^{th} bin is defined as $B_k \equiv [k\delta, (k\delta + \delta))$ and b_k is the bin count in bin B_k. Now, we add s of the adjacent bin counts $\{b_k\}$ to obtain the bin count for an ordinary histogram. The ASH estimate for x in B_o is the average of the heights of the s shifted histograms:

$$\frac{b_{1-s} + \cdots + b_0}{nh_b}, \frac{b_{2-s} + \cdots + b_0 + b_1}{nh_b}, \cdots, \frac{b_0 + \cdots + b_{s-1}}{nh_b}$$

A general expression for the unweighted ASH in equation (1) is

$$\hat{f}(x:s) = \frac{1}{s}\sum_{i=1-s}^{s-1} \frac{(s-|i|)b_{k+i}}{nh_b} = \frac{1}{nh_b}\sum_{i=1-s}^{s-1}(1 - \frac{|i|}{s})b_{k+i} \; for\, x \in B_k, \tag{2}$$

where b_k is a bin count, B_k is the k^{th} bin, s is the number of histograms, and h_b is a bin width.

Then weights on the bin counts in equation (2) may seem to take on the shape

[2] Consider a collection of m histograms, $\hat{f}_1, \hat{f}_2, \cdots, \hat{f}_n$. The bin width is the same, h. But bin origins differ: $t_0 = 0, \frac{h}{m}, \frac{2h}{m}, \cdots, \frac{(m-1)h}{m}$, respectively. Then the histograms are shifted according to different bin origins.
(See Scott, 2015, p.126, Figure 5.1 for example)

of an isosceles triangle with base $(-1, 1)$ or other shape.

Although the parameter s in the ASH is a nuisance parameter, it is more workable than the bin origin. The limiting behavior of the ASH as $s \to \infty$ has been extensively studied since 1950's. Finally the limit of ASH was obtained as a kernel estimator. With h_b and n fixed and s increasing, the effect of a single data point x_j on the ASH estimate $\hat{f}(x)$ can be isolated at a fixed point x. If $x \in B_k$ and $x_j \in B_{k+1}$, the index labeling of the bins changes as s increases. Then from equation (2), the influence of x_j on x is proportional to the weight

$$1 - \frac{|i|}{s} = 1 - \frac{|i| \cdot \delta}{s \cdot \delta} = 1 - \frac{|x - x_j|}{h_b} + O(\frac{\delta}{h_b}),\ if\ |x - x_j| < h_b,$$

where h_b is the bin width, s is the number of histograms, and $\delta = \dfrac{h_b}{s}$ at B_k .

Since x and x_j are in bins B_k and B_{k+i} respectively, the number of bins between these points is approximately i. Hence $|x - x_j| \approx |i| \cdot \delta$. Therefore, equation (2) may be rewritten:

$$\lim_{s\to\infty} \hat{f}(x : s) = \frac{1}{nh_b} \sum_{j=1}^{n} \left(1 - \frac{|x - x_j|}{h_b}\right) I_{[-1,1]} \left(\frac{x - x_j}{h_b}\right),$$

where the sum is not over the number of bins, but the number of data points n. If a kernel function $K(\cdot)$ is given as an isosceles triangle identy (cf. Scott, 2015, p.134),

$$\left(1 - |\frac{x - x_j}{h_b}|\right) I_{[-1,1]} \left(\frac{x - x_j}{h_b}\right) \text{ approaches } K \left(\frac{x - x_j}{h}\right).$$

That is, the limit of ASH may be written as

$$\lim_{s\to\infty} \hat{f}(x : s) = \frac{1}{nh} \sum_{i=1}^{n} K \left(\frac{x - x_i}{h}\right) = \hat{f}(x),$$

where h is a smoothing parameter(bandwidth).

The kernel estimate $\hat{f}(x)$ is a mixture density with n identical component densities (kernel functions) centered on the data points.

2.2 Properties of Kernel Estimator

Among many methods of density estimation, kernel density estimation will be discussed since "virtually all nonparametric algorithms are asymptotically kernel methods" (Scott, 2015, p.138). In fact, histograms use a rectangular kernel and the kernel density estimator (kernel estimator in short) is a generalized version of ASH. Thus the kernel estimator or the limiting ASH is written as

$$\begin{aligned} &\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) = \frac{1}{nh}\sum_{i=1}^{n} K(u), \\ &\text{where } h \text{ is the bandwidth } u = \frac{x-x_i}{h},\ K \text{ is the kernel function, and} \\ &K\left(\frac{x-x_i}{h}\right) \sim iid, \text{ symmetric around } x \end{aligned} \tag{3}$$

in a random sample $(X_1, \cdots, X_i, \cdots, X_n)$ of size n. Kernels may or may not be densities such as the normal density. The kernel estimator in equation (3) has the following properties.

Mean: $E\hat{f}(x) = EK_h = f(x) + O(h^2)$

Variance: $Var\hat{f}(x) = \frac{f(x)R(K)}{n \cdot h} - \frac{f(x)^2}{n} + O(\frac{h}{n})$, where $R(K) = \int_{-\infty}^{\infty} K(u)^2 du$

Bias: $Bias\hat{f}(x) = \frac{1}{2}h^2 f''(x)\sigma_K^2 + O(h^4)$, where σ_K^2 is $\int_{-\infty}^{\infty} u^2 K(u) du$.

These properties are proved in Appendix A.

2.3 Regression Estimator for Nonparametric Regression

Let us assume that there exists a smooth regression function $m(\cdot)$ for response y and the predictor x:

$$y_i = m(x_i) + \epsilon_i,\ for\, 1 \le i \le n,\ where\, \epsilon_i \sim f(\mu, \sigma^2(x)) \tag{4}$$

We often assume that f satisfies normality and homoscedasticity of the error term so that we may have a constant variance such that $\epsilon_i \sim N(\mu, \sigma^2)$. In this section, we will see how the kernel (density) is related to regression estimator. It is reasonable to view regression estimators as linear smoothers combining observed responses linearly. We will begin with a nonparametric smoother as an implementation of ASH estimate of kernel. Although there are some useful methods in nonparametric regression estimation (cf. Bowman and Azzalini, 1997; Györfi, Kohler, Krzyzak, and Walk, 2002; Scott, 2015), two kernel estimators will be focused in this research as they will be employed in estimation of conditional covariance matrices later. One is the local mean model and the other is the local linear model. The local mean model is a special form of the local linear model in that the former fits a local constant as a form of the linear model for data while the latter fits a full linear model. The typical local mean model is the one proposed by Nadaraya(1964, 1965) and Watson(1964). The local linear model has been pioneered by Fan and his colleagues (Fan, 1992, 1993; Fan and Gijbels, 1992, 1996).

2.3.1 The Nadaraya-Watson Estimator

The equation (4) can be considered in two distinct cases depending on what we assume regarding probabilistic structure in the data $(x_i, y_i) : 1 \leq i \leq n$. One is a fixed design assuming that data x_i are not random, but specifically chosen by the researcher. The other is a random design assuming that data come from a joint probability density function $f(x, y)$. The random design will be the focus in this section because we will use it in estimating conditional covariance matrices later on.

Let us define the theoretical regression as follows:

$$m(x) = E(Y|X = x) = \int yf(y|x)dy = \frac{\int yf(x, y)dy}{\int f(x, y)dy} \tag{5}$$

Based on equation (5) we can construct a nonparametric regression estimator using the kernel estimator we already discussed in the previous sections. Let

$f(x, y)$ be the unknown bivariate density. Then,

$$\begin{aligned}\hat{f}(x,y) &= \frac{1}{nh_xh_y}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right) K\left(\frac{y-y_i}{h_y}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} K_{h_x}K_{h_y} \text{ is kernel estimator of } f(X,Y).\end{aligned} \tag{6}$$

The kernel function K for $u_y = \dfrac{y-y_i}{h_y}$ makes the following assumptions (Hollander, Wolfe, and Chicken, 2014, p .618):

$K(u_y) \geq 0,\ -\infty < u_y < \infty,\ K(-u_y) = K(u_y),$
$\displaystyle\int_{-\infty}^{\infty} K(u_y)du_y = 1, \int_{-\infty}^{\infty} u_yK(u_y)du_y = 0.$

The denominator of equation (5) can be estimated by

$$\begin{aligned}\int_{-\infty}^{\infty} \hat{f}(x,y)\ dy &= \int_{-\infty}^{\infty} \frac{1}{n}\sum K_{h_x}K_{h_y}dy \text{ from equation (6)} \\ &= \frac{1}{n}\sum_{i=1}^{n} K_{h_x}\int_{-\infty}^{\infty}\frac{1}{h_y}K(\frac{y-y_i}{h_y})dy_i \\ &= \frac{1}{n}\sum_{i=1}^{n} K_{h_x}\int_{\infty}^{-\infty}\frac{1}{h_y}K(u_y)(-h_y)du_y \\ &= \frac{1}{n}\sum_{i=1}^{n} K_{h_x}\int_{-\infty}^{\infty}K(u_y)du_y = \frac{1}{n}\sum_{i=1}^{n} K_{h_x}\cdot 1 \\ &= \frac{1}{n}\sum_{i=1}^{n} K_{h_x} \\ &= \frac{1}{n}\sum K_h,\end{aligned}$$

where $\frac{y-y_i}{h_y} = u_y,\ \frac{du_y}{dy_i} = -\frac{1}{hy},\ Y_i = \infty \Rightarrow u_y = -\infty,\ Y_i = -\infty \Rightarrow u_y = \infty.$

Also, we can estimate the numerator of (5) by

$$\begin{aligned}
\int_{-\infty}^{\infty} y\hat{f}(x,y)\,dy &= \int_{-\infty}^{\infty} y_i \left(\frac{1}{n}\sum_{i=1}^{n} K_{h_x}K_{h_y}\right) dy_i \text{ from equation (6)} \\
&= \frac{1}{n}\sum_{i=1}^{n} K_{h_x} \int_{-\infty}^{\infty} y_i K\left(\frac{y-y_i}{h_y}\right) dy_i \\
&= \frac{1}{n}\sum_{i=1}^{n} K_{h_x} \int_{\infty}^{-\infty} y_i \frac{1}{h_y} K(u_y)(-h_y)du_y \\
&= \frac{1}{n}\sum_{i=1}^{n} K_{h_x} \int_{-\infty}^{\infty} y_i K(u_y)du_y \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i K_{h_x} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i K_h,
\end{aligned}$$

where $u_y = \frac{y-y_i}{h_y}$, $\frac{du_y}{dy_i} = \frac{-1}{h_y}$.

The resulting nonparametric kernel regression estimator for equation (5) was independently introduced by Nadaraya(1964, 1965) and Watson(1964) as

$$\begin{aligned}
\hat{m}(x) = \frac{\int_{-\infty}^{\infty} y\hat{f}(x,y)dy}{\int_{-\infty}^{\infty} \hat{f}(x,y)dy} &= \frac{\frac{1}{n}\sum_{i=1}^{n} y_i K_h}{\frac{1}{n}\sum_{i=1}^{n} K_h} \\
&= \frac{\sum_{i=1}^{n} y_i \frac{1}{h} K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x-x_i}{h}\right)} \\
&= \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}.
\end{aligned} \tag{7}$$

If we replace i with j in the denominator of equation (7) we have a little more convenient form:

$$\begin{aligned}
\hat{m}(x) &= \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)} \\
&= \sum_{i=1}^{n} y_i \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)} \\
&= \sum_{i=1}^{n} y_i g_i(x), \\
where\, g_i(x) &= \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}.
\end{aligned} \tag{8}$$

The weight $g_i(x)$ is determined by the following: choice of the kernel function K, bandwidth h on the values of a predictor x, and the distance between the observed data x_i and the point x. There are two notable features in this regression estimator. First, this estimator is a linear smoother in the observations of y_i, which is shared by many other nonparametric regression estimators using kernel methods. Second, the kernel estimate does not depend on what value of h_y is chosen as the smoothing parameter. This feature may be responsible for non-robustness of kernel regression (cf. Scott, 2019, p.243).

As we see in equation (8), the numerator and denominator are correlated random variables making the derivation of bias and variance very challenging. The bias, variance, and asymptotic MSE of Nadaraya-Watson estimator are given below(cf. Scott, 2015, pp.245-246):

$E\hat{m}(x) \approx m(x) + bias$,

$bias[\hat{m}(x)] = \frac{1}{2}h^2\sigma_K^2 \left[m''(x) + 2m'(x)\frac{f'(x)}{f(x)}\right]$,

$$Var[\hat{m}(x)] = \frac{R(K)\sigma^2}{nhf(x)}, \text{ where } R(K) = \int_{-\infty}^{\infty} K(u)^2 du, \text{ and}$$
$$AMSE[\hat{m}(x)] = \frac{R(K)\sigma^2}{nhf(x)} + \frac{1}{4}h^4\sigma_K^4 \left[m''(x) + 2m'(x)\frac{f'(x)}{f(x)} \right]^2,$$
where $\sigma_K^2 = \int_{-\infty}^{\infty} u^2 K(u)du$ and $\sigma_K^4 = \int_{-\infty}^{\infty} u^4 K(u)du$.

The local mean $\hat{m}(x) = \bar{y}$ obtained in Nadaraya-Watson estimator can be viewed as the best constant fit to data. An initial conceptualization can be given in

$$\bar{y} = \arg\, min_a \sum_{i=1}^{n} (y_i - a)^2, \tag{9}$$

where the arg $\min_a$ means that the constant $a = \bar{y}$ gives the minimum of the criterion sum, $\sum_{i=1}^{n}(y_i - a)^2$. The constant a is fit to data "locally" in that only those data (x_i, y_i) for $x_i \in (x - h, x + h)$ are included in the criterion sum. If we incorporate the kernel function $K(\frac{x - x_i}{h})$ as the weight in equation (9), the best local constant fit is

$$\hat{m}(x) = \arg\, min_a \sum_{i=1}^{n} K(\frac{x - x_i}{h})(y_i - a)^2 \tag{10}$$

Minimizing the weighted sum in the right hand side of equation (10) with respect to a leads to equation (8). The pointwise result in equation (8) is used to derive the entire regression function as follows:

$$\hat{m}(\cdot) = \arg\, min_{a(\cdot)} \int_{-\infty}^{\infty} \sum_{i=1}^{n} K(\frac{x - x_i}{h})[y_i - a(x)]^2 dx$$

as the integrand is minimized by $\hat{m}(x) = \hat{a}(x)$ in equation (10). This allows the Nadaraya-Watson estimator to be a special case of the local linear estimator or least square estimator(Györfi, Kohlen, Krzyzak, & Walk, 2002; Scott, 2015).

2.3.2 Local Linear Estimator

Let us assume that the local linear regression smoother $m(x)$ has the second derivative and $x_d = X_i - x$. By linearization of m at x, $m(x_d) \approx$

$m(x) + m'(x)(X_i - x) \equiv \alpha + \beta(X_i - x)$. Then, we have $\hat{m}(x) = \hat{\alpha}$ in a local linear regression problem. Drawing upon Stone(1997), $\hat{\alpha}$ and $\hat{\beta}$ are weighted least squares estimators minimizing

$$S = \sum_{i=1}^{n} [Y_i - \alpha - \beta(X_i - x)]^2 K\left(\frac{x - X_i}{h}\right)$$

given a random sample $(X_1, Y_1), \cdots (X_n, Y_n)$ from the distribution (X, Y).

Putting $K_i = K\left(\frac{x-X_i}{h}\right)$ and $C_i = X_i - x$, we have $S = \sum_{i=1}^{n} [Y_i - \alpha - \beta C_i]^2 K_i$. Its normal equations with $k_i = K(\frac{x-x_i}{h})$ and $c_i = x_i - x$ are

$$\hat{\alpha} \sum_{i=1}^{n} k_i + \hat{\beta} \sum_{i=1}^{n} k_i c_i = \sum_{i=1}^{n} k_i y_i \text{ and } \hat{\alpha} \sum_{i=1}^{n} k_i c_i + \hat{\beta} \sum_{i=1}^{n} k_i c_i^2 = \sum_{i=1}^{n} k_i c_i y_i$$

Then, by the Cramer Rule, $\hat{\alpha} = p/q$, where

$$p = \begin{vmatrix} \sum k_i y_i & \sum k_i c_i \\ \sum k_i c_i y_i & \sum k_i c_i^2 \end{vmatrix} \text{ and } q = \begin{vmatrix} \sum k_i & \sum k_i c_i \\ \sum k_i c_i & \sum k_i c_i^2. \end{vmatrix}$$

As a result

$$\hat{\alpha} = \frac{\sum k_i y_i \sum k_i c_i^2 - \sum k_i c_i \sum k_i c_i y_i}{\sum k_i \sum k_i c_i^2 - (\sum k_i c_i)^2} \tag{11}$$

Putting $d_i = x - x_i$, $s_1 = \sum k_i d_i$ and $s_2 = \sum k_i d_i^2$, the numerator of equation (11) is

$$\begin{aligned}
&\sum k_i y_i \sum k_i d_i^2 - \sum k_i d_i \sum k_i d_i y_i \\
&= (\sum k_i y_i) s_2 - (\sum k_i d_i y_i) s_1 \\
&= \sum k_i s_2 y_i - \sum k_i d_i s_1 y_i \\
&= \sum (k_i s_2 - k_i d_i s_1) y_i \\
&= \sum k_i (s_2 - d_i s_1) y_i \\
&= \sum w_i y_i, \text{ where } w_i = k_i (s_2 - d_i s_1).
\end{aligned}$$

And the denominator of equation (11) is

$$\begin{aligned}
&\sum k_i \sum k_i d_i^2 - (\sum k_i d_i)^2 \\
&= \sum k_i s_2 - \sum k_i d_i s_1 \\
&= \sum k_i (s_2 - d_i s_1) \\
&= \sum w_i .
\end{aligned}$$

Then, $\hat{m}(x) = \hat{\alpha} = \dfrac{\sum w_i y_i}{\sum w_j}$.

Since $\dfrac{w_i}{\sum w_j}$ is a weight, $\hat{m}(x) = \sum y_i z_i(x)$, where $z_i(x)$ is the weight. The $z_i(x)$ becomes a probability weight function if $z_i(x) \geq 0$ and $\sum_{i=1}^{n} z_i(x) = 1$ (Stone, 1977). The $z_i(x) = z_i(x, X_1, \cdots, X_n), 1 \leq i \leq n$, gives more weights on the values of y_i for which X_i is close to x and less weights on the values of y_i for which X_i is far from x .

The properties of $\hat{m}(x)$ are as follows.

$E[\hat{m}(x)] \approx m(x) + \dfrac{h^2}{2}\sigma_K^2 m''(x)$,

$Var[\hat{m}(x)] \approx \dfrac{\sigma^2 R(K)}{nhf(x)}$, where $R(K) = \displaystyle\int_{-\infty}^{\infty} K(u)^2 du$,

and $MSE = \dfrac{1}{4}[\sigma_K^2 m''(x)]^2 h^4 + \dfrac{R(K)}{nh} \cdot \dfrac{\sigma^2}{f(x)} + o(h^4 + \dfrac{1}{nh})$.

(cf: Bowman & Azzalini, 1996, p.70; Fan, 1992, p.1000)

2.3.3 Other Kernel Estimators

Two more kernel estimators are worthy of mention. Priestly and Chao(1972) introuced incorporation of the distance between adjacent observed points x_i. The idea is to add extra weight in addition to the kernel function:

$\hat{m}(x) = \displaystyle\sum_{i=1}^{n} y_i \frac{(x_i - x_{i-1})}{h} K(\frac{x - x_i}{h})$, where $x_{i-1} \leq x_i$.

On the other hand, Gasser and Müller(1979) showed an integration of the kernel function in a small neighborhood of x_i.

$\hat{m}(x) = \displaystyle\sum_{i=1}^{n} y_i \frac{1}{h} \int_{t_i - 1}^{t_i} K(\frac{x - x_i}{h}) dt$,

where $t_0 = -\infty, t_i = \frac{1}{2}(x_i + x_{i+1}), t_n = \infty$, and x_i denotes the i^{th} largest value

of the observed predictor values.

Priestly and Chao's estimator showed a larger mean integrated squared error than the Nadaraya-Watson estimator and appeared to be not consistent at the end points(Benedetti, 1974). Gasser and Müller's estimator has larger variance than the Nadaraya-Watson estimator, but smaller bias (Fan, 1992). After comparing seven different kernel estimators, Benedetti(1974) concluded that, at least asymptotically, there is little difference in the choice of kernel. However, Fan(1992) showed that local linear estimator(LLE) overcomes three shortcomings of Nadaraya-Watson and Gasser-Müller estimators. First, the LLE can be adapted to both random and fixed designs. Second, LLE does not require equal spacing in X_i enabling it to adapt to both highly clustered and nearly uniform designs. Third and finally, LLE's behavior near the edges of the data space is superior to other estimators. Also, Fan(1992) showed that LLE is not only superior to Nadaraya-Watson and Gasser-Müller estimators, but also it is the best among all linear smoothers including those produced by kernel, orthogonal series, and spline methods.

2.4 Bandwidth Selection

The smoothing parameter in nonparametric kernel estimation is called bandwidth. Bandwidth is one of the three factors determining the effectiveness of nonparametric kernel regression: the three factors are kernel function, distance $(x-x_i)$, and bandwidth h. It has been well studied that different choice of kernel functions makes little difference in estimation(e.g., Benedetti, 1974), and weighted distance is determined by the kernel function and choice of bandwidth. So, the bandwidth h is the critical factor that determines the width of the kernel function and the degree of smoothing (cf. Bowman & Azzalini, 1996, p.49).

It is inevitable that there is a tradeoff between large h and small h in that the former causes large bias and small variance, and the latter causes small bias and large variance. As the criterion to evaluate effectiveness of h, mean squared error(MSE) is used: $MSE(h) = E[\hat{m}(x) - m(x)]^2$. Also, for asymptotic analysis, mean integrated squared error(MISE) can be considered:

$MISE(h) = \int E[\hat{m}(x) - m(x)]^2 f(x)\ dy$. Cross-validation provides a means of selecting h by minimizing an estimate of MISE over different values of h.

Two major approaches to bandwidth selection are cross-validation and plug-in. The former is so popular that it is called in different ways: n-fold cross-validation (Györfi, Kohler, Krzyzak, & Walk, 2002), ordinary cross-validation(Wood, 2017), or cross-validation(Bowman & Azzalini, 1996; Scott, 2015).

The cross-validation criterion is

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{m}_{-i}(x_i)]^2, \tag{12}$$

where the subscript $-i$ means leaving out the i^{th} data (x_i, y_i) in computing $\hat{m}(x_i)$. The optimal h is the one minimizing $CV(h)$ among possible h values.

As an alternative to cross-validation approach, plug-in estimator of the optimal h is defined as

$$h_{opt} = \left[\frac{R(K)\sigma^2}{n\int\{m''(x)\}^2 f(x)dx}\right]^{1/5}.$$

There is a difficulty in estimating the second derivative $m''(x)$ of the regression function. Gasser, Kneip, and Köhler(1991) provide a formula for h_{opt} suitable to Gasser-Müller estimator. Ruppert, Sheather, and Wand(1995) provides the same thing for the local linear estimator. Although a plug-in approach is very promising, cross-validation will be employed in this research as it can be applied to a wide variety of settings (Bowman and Azzalini, 1997).

2.5 A Simple Practice of Nonparametric Kernel Regression

A sine function was used to develop 50 sample data points as in Bowman and Azzalini(1996, p.79). Also a nonparametric kernel regression program is prepared in R language. As the kernel function, normal density was employed: $\frac{1}{\sqrt{2\pi}h}e^{-\frac{1}{2}(\frac{x-X_i}{h})^2}$. The local linear model was employed as the regression estimator. The optimal bandwidth was selected by finding the h value minimizing

the cross-validation criterion in equation (12). To evaluate the criterion, 32 values of h were used as was done in Bowman and Azzalini(1996, p.79). Once the optimal h was selected, it was used to compute regression estimates $\hat{m}(x)$.

As a result of running the program, Figure 1 was obtained for CV criterion score vs h values. Figure 1 shows that the CV score is lowest around $h = 0.05$. The output showed $hcv.opt = 0.0561396$ which is automatically used to produce $\hat{m}(x)$. The Figure 2 shows the true sine curve (full) and regression curve (dotted).

Figure 1. CV criterion score vs h

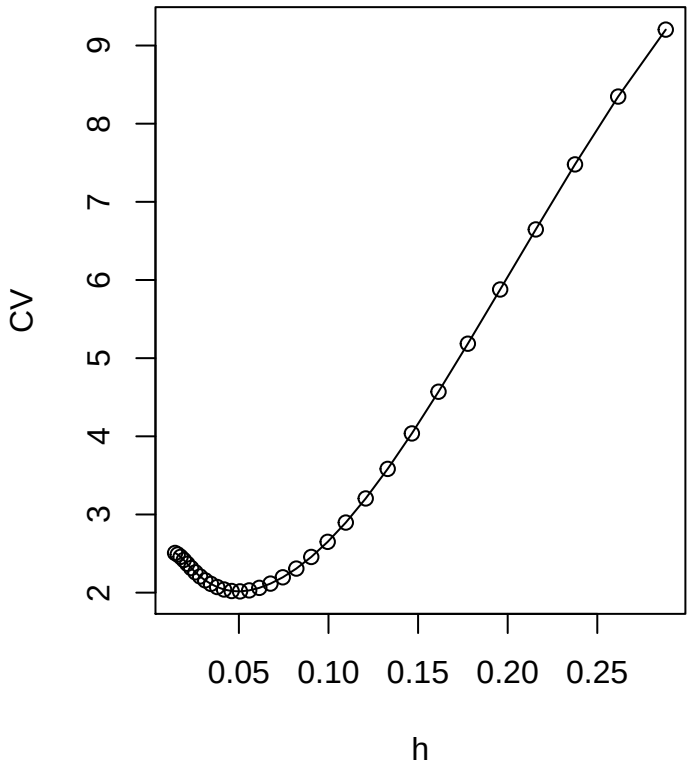
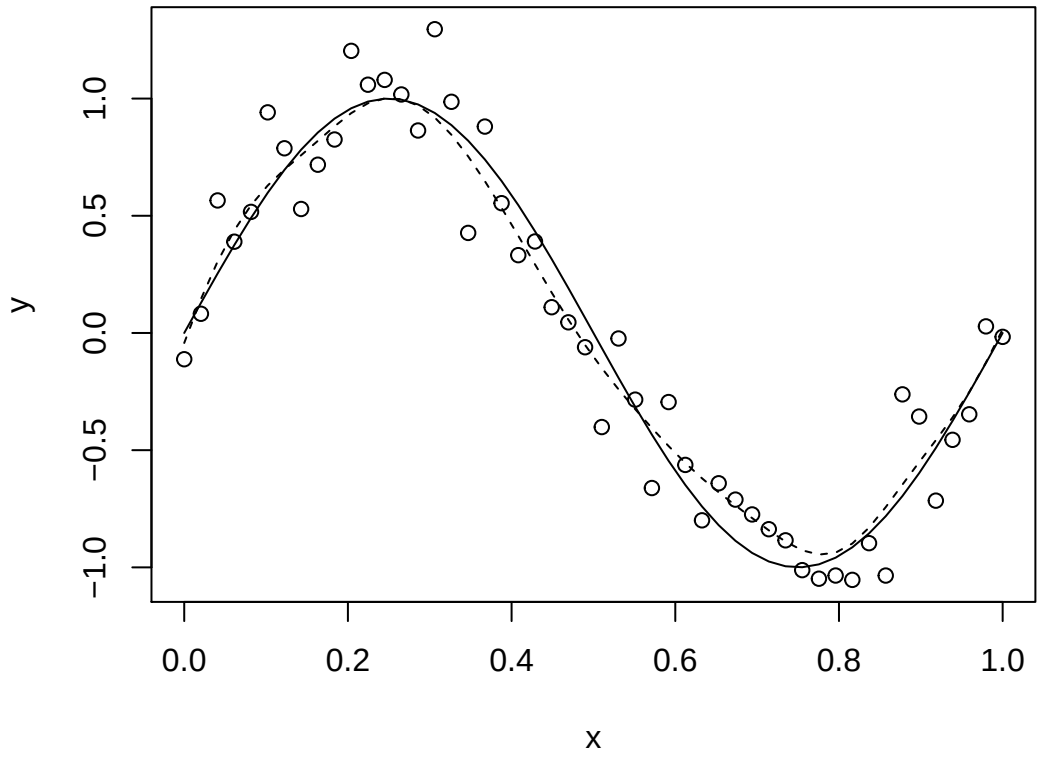


Figure 2. True (full) and Regression (dotted) Curves



3 Estimator of Conditional Covariance Matrices

So far we have considered nonparametric estimation of means employing a local mean (local constant) model and local linear model with the former being a special case of the latter. Now we can think of estimation of conditional variances and conditional covariances as components of a conditional covariance matrix. Nonparametric conditional variances have been studied earlier. Ruppert, Wand, Holst, and Hössjer (1997) proposed local polynomial variance-function estimation based on linear smoothing of squared residuals. Fan and Yao (1998) proposed a similar idea that can be applied to nonlinear

time series data. However, they did not explicitly recognize that local linearity or higher-degree polynomials would threaten the positive-definiteness of variances in their estimation. To avoid this problem Yu and Jones (2004) proposed local polynomial models taking log of conditional variances.

In estimating nonparametric conditional covariance matrices, Yin, Geng, Li, and Wang (2010) employed a local constant model instead of a local linear or higher-degree model to ensure positive-definiteness of their nonparametric kernel estimator. Since we are interested in estimation of conditional covariance matrices with conditional variances being a part of them, Yin et al's work will be discussed in detail.

3.1 Yin et al's (2010) Nonparametric Conditional Covariance Model

Suppose that we have two vectors of variables V and X. The model of the conditional covariance of V given X as an index random vector is $cov(V|X) = \sum(X)$. Random variables are:

$V = (V_1, \cdots, V_p)^T \in \Re^p$ for variables of interest;

$X = (X_1, \cdots, X_q)^T \in \Re^q$ for conditioning variable or index random vector. Due to the curse of dimensionality, q=1 only is chosen for X here.

Assumptions are:

(1) $V \sim (m(x),\ \sum(x))$ conditional on $X = x$,

where $m(x) = \{m_1(x), \cdots, m_k(x), \cdots, m_p(x)\}^T$ and $\sum(x) = \{\sigma_{j_1 j_2}(x)\}$;

(2) Both $m(x)$ and $\sum(x)$ are unknown but smooth functions of x.

The $m(x)$ can be estimated using local constant or local linear model. However, Yin et al developed a Nadaraya-Watson(NW) kernel estimator for both $m(x)$ and $\sum(x)$ to estimate $\sum(x)$ consistently.

Let $K(x)$ be a symmetric kernel density function, and $K_h(x) = \frac{1}{h}K(\frac{x}{h})$ be the scaled kernel function with a bandwidth $h > 0$. To motivate estimation, it is temporarily assumed that given X, V follows a normal distribution, although the conditional normality assumption is unnecessary. Suppose (V_i, x_i) with

$V_i = \begin{pmatrix} V_{i1} \\ \vdots \\ V_{ip} \end{pmatrix}$, $i = 1, \cdots, n$ is a random sample from the population of (V, X).

Due to the normality assumption, $V_i = m(x) + \underline{\varepsilon_i}$, $\underline{\varepsilon_i} \sim N_p(\underline{0}, \sum(x)$,

where $f(v_k) = \frac{1}{\sqrt{2\pi}\sigma(x)} exp[-\frac{1}{2}(\frac{v_{ki} - m_k(x)}{\sigma(x)})^2]$, $i = 1, \cdots, n$.

Then the likelihood function L follows:

$$L = f(V) = \frac{exp[-\frac{1}{2}(V - m(x))^T \sum^{-1}(x)(V - m(x))]}{(2\pi)^{np/2} |\sum(x)|^{n/2}}.$$

Let Λ be the natural logarithm of L:

$$\Lambda(m(x), \sum(x)) = \log L = -\frac{1}{2}(V - m(x))^T \sum{}^{-1}(x)(V - m(x)) - \frac{np}{2}\log 2\pi - \frac{n}{2}\log|\sum(x)| \tag{13}$$

Note that maximizing L is equivalent to maximizing Λ or minimizing $-\Lambda$

$= \frac{1}{2}(V - m(x))^T \sum^{-1}(x)(V - m(x)) + \frac{np}{2}\log 2\pi + \frac{n}{2}\log|\sum(x)|$,

where $\frac{np}{2}\log 2\pi = \text{constant}$, $\frac{n}{2}\log|\sum(x)| = -\frac{n}{2}\log|\sum^{-1}(x)|$.

Let $A = \frac{1}{2}(V - m(x))^T \sum^{-1}(x)(V - m(x)) - \frac{n}{2}\log|\sum^{-1}(x)|$,

$B = (V - m(x))^T \sum^{-1}(x)(V - m(x)) - n\log|\sum^{-1}(x)|$, and

$C = \sum_{i=1}^{n}[(V_i - m(x))^T \sum^{-1}(x)(V_i - m(x)) - \log|\sum^{-1}(x)|]$.

Then minimizing A is equivalent to minimizing B, which is equivalent to minimizing C. If we add $K_h(x_i - x)$ as the weight, the kernel method is to minimize "weighted C" $= C'$:

$C' = \frac{1}{n}\sum_{i=1}^{n}[(V_i - m(x))^T \sum^{-1}(x)(V_i - m(x)) - \log|\sum^{-1}(x)|]K_h(x_i - x)$, where x is an arbitrary point in the support of X. This C' is the equation (2.1)

of Yin et al (2010).

To derive nonparametric kernel estimators of mean vector and covariance matrix, we will begin with equation (13). Given $\Lambda(m(x), \sum(x)) = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\sum(x)| - \frac{1}{2}\sum_{i=1}^{n}[(V_i - m(x))^T \sum^{-1}(x)(V_i - m(x))]$,

$$
\begin{aligned}
d\Lambda &= -\frac{n}{2} d\log|\sum(x)| + \frac{1}{2}\sum_{i=1}^{n}(dm(x))^T \sum{}^{-1}(x)(V_i - m(x)) \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}(V_i - m(x))^T (d\sum{}^{-1}(x))(V_i - m(x)) \\
&\quad + \frac{1}{2}\sum_{i=1}^{n}[(V_i - m(x))^T \sum{}^{-1}(x) dm(x)] \\
&= -\frac{n}{2} d\log|\sum(x)| - \frac{1}{2}\sum(V_i - m(x))^T (d\sum{}^{-1}(x))(V_i - m(x)) \\
&\quad + \sum_{i=1}^{n}[(V_i - m(x))^T \sum{}^{-1}(x) dm(x)] \\
&= -\frac{n}{2} tr[\sum{}^{-1}(x) d\sum(x) + S d\sum{}^{-1}(x)] + \sum_{i=1}^{n}[(V_i - m(x))^T \sum{}^{-1}(x) dm(x)], \\
&\quad \text{where } S = S(m(x)) = \frac{1}{n}\sum_{i=1}^{n}(V_i - m(x))(V_i - m(x))^T.
\end{aligned}
$$

Let $\hat{m}(x)$ and $\hat{\sum}(x)$ be the MLEs of $m(x)$ and $\sum(x)$ respectively.

Let $\hat{S} = S(\hat{m}(x))$ and set $d\Lambda = 0$, then we have

$tr[\hat{\sum}^{-1}(x) - \hat{\sum}^{-1}(x)\hat{S}\hat{\sum}^{-1}(x)]d\sum(x) = 0$ for all $d\sum(x)$ and

$\sum_{i=1}^{n}(V_i - \hat{m}(x))^T \hat{\sum}^{-1}(x) dm(x) = 0$ for all $dm(x)$.

Thus, $\hat{\sum}^{-1}(x) = \hat{\sum}^{-1}(x)\hat{S}\hat{\sum}^{-1}(x)$ and $\sum_{i=1}^{n}(V_i - \hat{m}(x)) = 0$.

Hence, the MLEs are:

$$\hat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} V_i, \tag{14}$$

and

$$\hat{\sum}(x) = \frac{1}{n}\sum_{i=1}^{n}(V_i - \hat{m}(x))(V_i - \hat{m}(x))^T. \tag{15}$$

However, considering that we use a NW kernel method for nonparametric estimation, we need to substitute g_i for $\frac{1}{n}$ in equation (14) and equation (15), where $g_i = g_i(x) = K(\frac{x_i - x}{h}) / \sum_{j=1}^{n} K(\frac{x_j - x}{h})$ from equation (8). So, we have NW kernel estimators of $m(x)$ and $\sum(x)$ as follows:

$$\hat{m}(x) = \sum_{i=1}^{n} g_i V_i = \frac{\sum_{i=1}^{n} K_h(x_i - x)V_i}{\sum_{j=1}^{n} K_h(x_j - x)} \tag{16}$$

and

$$\begin{aligned} \hat{\sum}(x) &= \sum_{i=1}^{n} g_i (V_i - \hat{m})(V_i - \hat{m})^T \\ &= \frac{\sum_{i=1}^{n} K_h(x_i - x)(V_i - \hat{m}(x_i))(V_i - \hat{m}(x_i))^T}{\sum_{j=1}^{n} K_h(x_j - x)} \end{aligned} \tag{17}$$

The $\hat{m}(x)$ is identical with the Nadaraya-Watson estimator in the equation (8) for nonparametric regression which was obtained without the normality assumption. Estimation of $m(x)$ and $\sum(x)$ can be done separately. Moreover, we can use different bandwidths for different components of the conditional mean vector and also the conditional covariance matrix. However, Yin et al. suggest using the same bandwidth for all elements of covariance matrix to ensure positive-definiteness. For simplicity, Yin et al. used the following log likelihood type cross-validatory criterion to select the bandwidth for estimation of a conditional covariance matrix:

$$CV_{\sum}(h) = \frac{1}{n}\sum_{i=1}^{n} [\{V_i - \hat{m}(x_i)\}^T \hat{\sum}_{(-i)}^{-1}(x_i)\{V_i - \hat{m}(x_i)\} - \log(|\hat{\sum}_{(-i)}^{-1}(x_i)|)],$$

where $\hat{\sum}_{(-i)}$ is the estimate computed according to $\hat{\sum}(x)$ but without the i^{th} observation. The bandwidth for the $\hat{\sum}(x)$ is determined by minimizing $CV_{\sum}(h)$.

3.2 Theoretical Properties of $\hat{\sum}(x)$

Since the mean estimator has been already studied in the section of NW regression estimator, we will study only the asymptotic properties of the conditional covariance estimator. Under the Gaussian kernel which is widely employed, Yin et al. (2010) employ the following five regularity conditions to facilitate the study of asymptotic properties.

C.1 (The density of the index variable)

The range of values for the probability density function of X, denoted by $f(x)$, is closed and bounded: the range where $f(x) > 0$ looks like $[a, b]$ with a and b real numbers. The $f(x)$ is twice differentiable and its derivatives are continuous functions of X.

C.2 (The moment requirement)

For any $1 \leq j_1,\ j_2 \leq p$, there exists a constant $\delta \in [0, 1)$ such that $sup_x\ E\{|V_{j_1}(x)V_{j_2}(x)|\}^{2+\delta} < \infty$.

C.3 (Smoothness of the conditional mean)

The conditional mean $m_j(\cdot)$ is twice differentiable and its derivatives are continuous functions of X.

C.4 (Smoothness of the conditional variance)

The conditional variance is twice differentiable and its derivatives are continuous functions of X.

C.5 (The bandwidth)

$h \to 0$ and $nh^5 \to c > 0$ for some $c > 0$.

Denote $\hat{\sum}(x) = \{\hat{\sigma}_{j_1 j_2}(x)\}$ with $1 \leq j_1,\ j_2 \leq p$. For an arbirary function $q(x)$, we use $\dot{q}(x)$ and $\ddot{q}(x)$ to denote its first and second order derivatives, respectively. Also, it is defined that $\nu_0 = \int_{-\infty}^{\infty} K^2(x)dx$ and $\mu_2 = \int_{-\infty}^{\infty} x^2 K(x)dx$. Then Yin et al. show the asymptotic behavior of the covariance estimator $\hat{\sigma}_{j_1 j_2}(x)$ in two theorems.

Theorem1: $\sqrt{nh}\{\hat{\sigma}_{j_1j_2}(x) - \sigma_{j_1j_2}(x) - \theta_n\} \underrightarrow{D} N(0, f^{-1}(x)\nu_0 w_{j_1j_2}(x))$,
where

$$\theta_n = \frac{h^2\mu_2}{2}\{\ddot{\sigma}_{j_1j_2}(x) + 2\dot{\sigma}_{j_1j_2}(x)\frac{\dot{f}(x)}{f(x)}\}, \tag{18}$$

$\epsilon_{j_1j_2}(i) = \{V_{ij_1} - m_{j_1}(x_i)\}\{V_{ij_2} - m_{j_2}(x_i)\} - \sigma_{j_1j_2}(x_i)$ and

$w_{j_1j_2}(x_i) \underset{=}{\triangle} \text{var}(\epsilon_{j_1j_2}(i)|x_i)$.

This theorem shows that the variance of $\hat{\sigma}_{j_1j_2}(x)$ is of the order $\frac{1}{nh}$ and its bias is of the order h^2. So, Yin et al.'s covariance estimator is not unbiased, although its variance vanishes when n goes to infinity.

To evaluate the global convergence of $\hat{\sum}(x)$, Yin et al. suggest two loss functions that are explained in Muirhead (1982).

Stein loss $\Delta_1(x) = E\{tr\{\sum^{-1}(x)\hat{\sum}(x)\} - \log|\sum^{-1}(x)\hat{\sum}(x)|\} - p$

Quadratic loss $\Delta_2(x) = E[tr\{(\hat{\sum}(x)\sum^{-1}(x) - I)^2\}$

As the value of loss function gets smaller, the better the estimated covariance matrix converges to the population covariance matrix. The relationship between the two loss functions is given in Yin et al.'s theorem 2.

Theorem2: $\Delta_1(x) = 0.5\Delta_2(x)\{1 + o(1)\}$.

This theorem implies that choice between the Stein loss and the quadratic loss is not important as they differ only by just a constant and some negligible term asymptotically.

4 Simulation Study

A simulation study was conducted following Yin et al (2010) to see how the covariance estimator works. For the simulation, the following nonparametric conditional covariance model was employed for five study variables $(V_1, \cdots, V_5)$ and one conditioning or random index variable(X).

The mean vector was $m(x) = (6x, 10\cos(x), 25\sin(x), 20\exp(x), 30x - 10)^T \in \Re^5$ and the covariance matrix was $\sum(x) = C(x)C^T(x)$, where

$$C(x) = \begin{bmatrix} 2\cos(x) & 0 & 0 & 0 & 0 \\ \frac{3}{2}\sin(x) & 4\cos(x) & 0 & 0 & 0 \\ 2\sin(x) & \frac{5}{2}\sin(x) & 6\cos(x) & 0 & 0 \\ \frac{5}{2}\sin(x) & 3\sin(x) & \frac{7}{2}\sin(x) & 8\cos(x) & 0 \\ 3\sin(x) & \frac{7}{2}\sin(x) & 4\sin(x) & \frac{9}{2}\sin(x) & 10\cos(x) \end{bmatrix}.$$

For samples of different sizes (n=200, 400, and 800), 200 replications were analyzed. The number of evaluation points (points where kernel density was evaluated) on the index variable was $ne = 50$. The procedures of developing sample data are given in Appendix B.

Based on sample data, a cross-validated bandwidth for each of the five study variables is generated following the local constant model in kernel estimation and regression estimation. Mean estimates were obtained at 50 evaluation points using the cross-validated bandwidths of the study variables.

At the stage of estimating a conditional covariance matrix at each evaluation point, a cross-validated optimal bandwidth was computed following the equation of $CV_{\sum}(h)$ in Yin et al (2010, Section 2.3). This optimal bandwidth was applied to all five study variables. As there were 50 evaluation points(ne) 50 covariance matrices were estimated in each of 200 replications(rep).

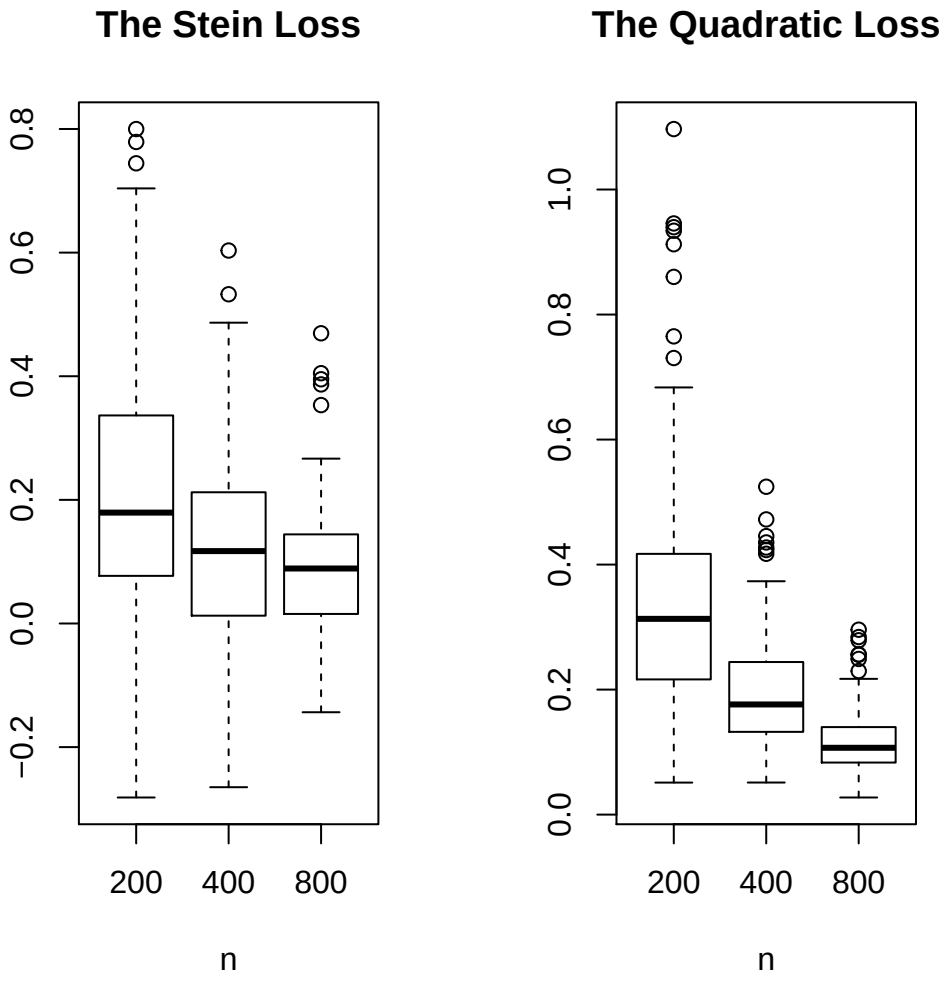
As the measure of estimation errors, global convergence indices such as Stein loss function value and quadratic loss function value were obtained by comparing the population conditional covariance matrix and each estimated conditional covariance matrix at each of 50 evaluation points. Summary of median values across 50 evaluation points for Stein loss function and quadratic loss function are given in Table 1 and Figure 3.

Figure 3 is similar to Figure 3.2 in the appendix of Yin et al (2010).

In general, the loss function values decrease as the sample sizes increase. Quadratic loss function shows higher values than those of Stein loss function. However, the difference is getting smaller and smaller as the sample size gets larger. The asymptotic equivalence between the two loss function values was indicated in the theorem2 of Yin et al.

In computation of Stein's loss function, Ledoit and Wolf (2017) suggests dividing function values by the number of study variables for normalization.

Figure 3. Box−plot of Estimation Errors for Simulation Study



However, in order to make my results comparable to Yin et al., I followed the equation (2.6) of Yin et al.

We do not know what level of the loss function values is acceptable in practice. Therefore to examine the behavior of loss function in estimation of conditional covariance matrices, there is a need for a large scale Monte-Carlo study.

Table 1: Summary of Simulation

	n=200, ne=50, rep=200					
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Stein loss	-0.2815	0.0775	0.1794	0.2078	0.3365	0.7999
Quadratic loss	0.0512	0.2163	0.3132	0.3435	0.4159	1.0967
	n=400, ne=50, rep=200					
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Stein loss	-0.2649	0.0127	0.1171	0.1209	0.2122	0.6034
Quadratic loss	0.0513	0.1330	0.1762	0.1962	0.2441	0.5245
	n=800, ne=50, rep=200					
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Stein loss	-0.1436	0.0156	0.0890	0.0839	0.1438	0.4695
Quadratic loss	0.0274	0.0834	0.1069	0.1161	0.1400	0.2958

5 Application to Boston Housing Data

To demonstrate an application of the covariance estimator to real data, I used a dataset collected by the U.S. Census Service concerning housing in the areas of Boston, Mass. Yin et al. (2010) analyzed the same dataset for application. The data were originally published by Harrison and Rubinfeld (1978). There are two versions for this dataset: original dataset and corrected dataset. The latter was downloaded from the StatLib archive (http://lib.stat.cmu.edu/datasets/boston-corrected.txt). The name of this dataset is "boston".

There are 506 observations (cases) from 91 towns (5.56 cases per town) in boston. Although there are many variables in the dataset, the following five variables were selected for analysis as in Yin et al. for comparison purpose.

CRIM - per capita crime rate by case
TAX - full-value property-tax rate per $10,000 by town
PRATIO - pupil-teacher ratio by town
MEDV - median value of owner-occupied homes in $1,000's by case
LSTAT - % lower status of the population by case

Identical values of TAX are assigned if the cases are collected from the same town. So is the case with PRATIO. Out of the five variables, the first four are the study variables of interest and LSTAT is the conditioning or random index variable. Our goal is to examine how the correlation structure of the four study variables vary along the values of the index variable.

As the local mean (constant) model was employed for kernel estimation, there might be large bias at the boundary (Yin et al, 2010). To alleviate this potential problem, we followed Yin et al who defined the index variable as the rank of LSTAT devided by total sample size, which forces the index variable to follow a uniform distribution with [0,1]. We will refer to this transformed LSTAT as t_LSTAT. All four study variables were standardized to have zero mean and unit variance as in Yin et al.

Two studies were conducted in this application. First, the dataset was randomly split into halves with equal sample size (training data and testing data) and they were used to confirm whether $\sum(x)$ varies along t_LSTAT. Second, the whole dataset (n=506) was analyzed and pointwise 90% confidence intervals around pairwise nonparametric correlation coefficients are computed.

5.1 Confirmation of Conditionality of Covariances

To examine how the covariance matrix of the four study variables varies along the index variable t_LSAT, the dataset boston was randomly split into training data and testing data with equal size. Two analyses were performed: One for examining the predictive performance of covariance estimator, the other for showing the poor performance of a null model based on the hypothesis that the covariances do not vary along the index variable.

To examine predictive performance of the covariance estimator, we need to estimate conditional means $\hat{m}(x)$ as in equation (16) and conditional covariance matrix $\hat{\sum}(x)$ as in equation (17) of testing data.

The predictive performance of the covariance estimator is evaluated by computing the forecasting error by the following out-of-sample loss measure:
$\Delta_{out} = \frac{1}{n^*} \sum_{i=1}^{n^*} [Y_t^T \hat{\sum}^{-1}(x_i^*) Y_t - \log(|\hat{\sum}^{-1}(x_i^*)|)]$,
where * stands for testing data, $n^* = n/2$, and $Y_t = V_i^* - \hat{m}(x_i^*)$ for testing

data (Yin et al., 2010). If both $m(x)$ and $\sum(x)$ are estimated accurately, we expect to obtain a good out-of-sample fit (small values of Δ_{out}) by treating their estimates as if they were parameters. From 200 replications of such an experiment, I obtained $\Delta_{out} = 0.5096$ (0.5820 in Yin et al., 2010).

To examine poor performance of a null model and to confirm conditionality of $\sum(x)$ on x, the same experiment was replicated another 200 times but with $\hat{\sum}(x)$ in equation (17) by

$$\tilde{\sum} = \frac{1}{n^*} \sum_{i=1}^{n^*} Y_t Y_t^T, \tag{19}$$

where $n^* = n/2, Y_t = V_i^* - \hat{m}(x_i^*)$ in testing data. The covariance matrix in equation (19) represents a null hypothesis that the covariances do not vary along the index variable. We expect that the resulting Δ_{out} would be much greater than 0.5096 of the nonparametric covariance model representing a poor hypothesis. The resulting Δ_{out} was 1.5267 (1.1056 in Yin et al., 2010) which was much larger than 0.5096 of the nonparametric covariance model, confirming our expectation. Then, we can conclude that nonparametric covariance model fits much better than the null model. That is, the covariance structure of the four social economic variables tends to vary along the proportion of the lower status population. This is the same finding as in Yin et al.

5.2 Interval Estimation of Conditional Correlation Coefficients

Based on the mean estimates of nonparametric covariances and bootstrapping, interval estimates of correlations (standardized form of covariances) are obtained. First, the sample correlation coefficients (Table 2) were obtained from the whole 506 observations of boston dataset. Interval estimates of conditional correlation coefficients will be compared to these unconditional correlation coefficients.

I computed bootstrap standard errors for the correlation estimates that will be used to construct interval estimates of correlation coefficients. Sub-

Table 2: Sample Correlation Matrix

	CRIM	**TAX**	**PTRATIO**	**MEDV**
CRIM	1			
TAX	0.5828	1		
PTRATIO	0.2899	0.4609	1	
MEDV	-0.3883	-0.4685	-0.5078	1

sequently, I plot the regression line, the upper limit line, and the lower limit line of correlation coefficients with 90% of confidence level. These are graphs of interval estimates of correlation coefficients conditional on the index variable, t_LSTAT. The pointwise 90% confidence intervals are given by $\hat{\sigma}_{j_1 j_2}(x) \pm 1.64\hat{SE}\{\hat{\sigma}_{j_1 j_2}(x)\}$, where x=t_LSTAT. The standard error estimate $\hat{SE}\{\hat{\sigma}_{j_1 j_2}(u)\}$ is computed based on 200 replications of bootstrapping. Figure 4 shows interval estimates of six correlation coefficients among the four study variables, and shows similar patterns to those in Figure 3.1 of Yin et al.

In Figure 4, index variable is the t_LSTAT and the correlations among four standardized study variables are plotted as the function of t_LSTAT. The conditionality of each correlation coefficient is examined along the values of t_LSTAT in comparison with the corresponding sample correlation coefficient.

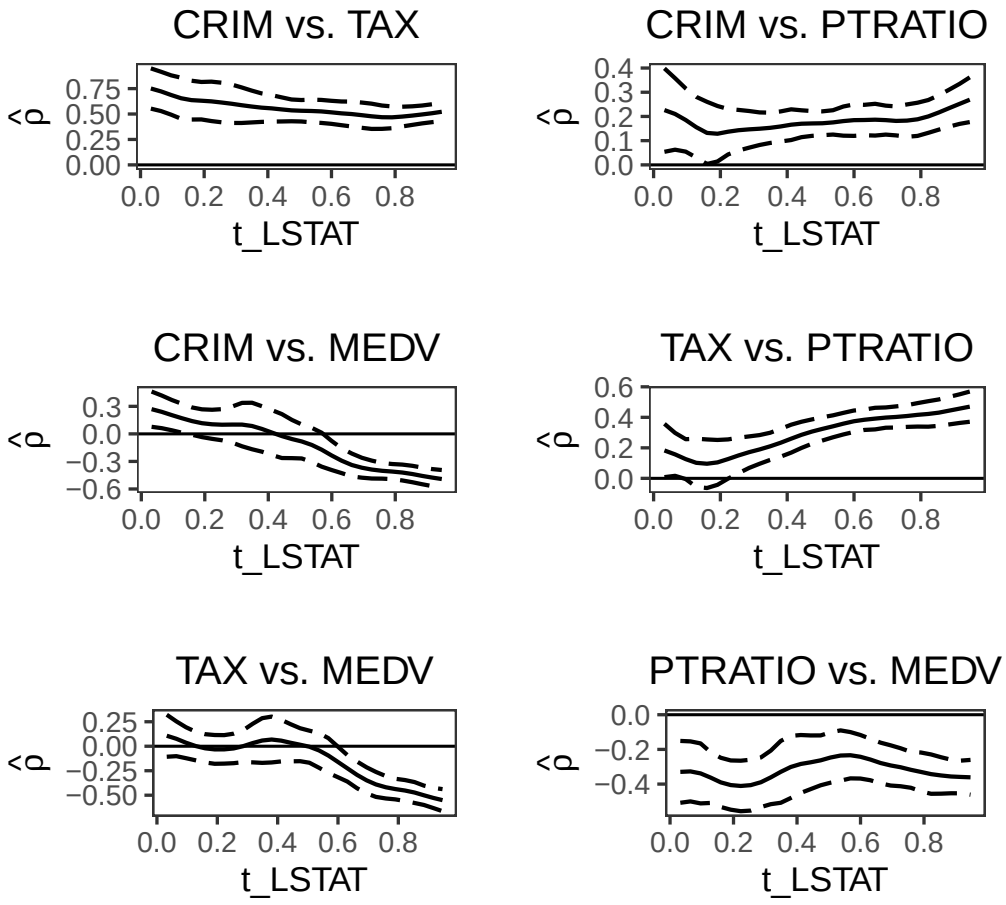
CRIM vs TAX: The sample correlation is 0.5828. The conditional correlations in Figure 4 are not extremely deviant from the sample correlation.

CRIM vs PTRATIO: The sample correlation is 0.2899. The conditional correlations in Figure 4 are not extremely deviant from the sample correlation.

Regarding the above two correlations, the difference between unconditional and conditional correlations is not particularly conspicuous as we observe the relatively stable pattern of interval estimates of conditional correlation coefficients. However, we will see drastically different patterns of interval estimates for some other conditional correlations.

CRIM vs MEDV: The sample correlation is -0.3883. However, Figure 4 shows a decreasing trend of the correlation as the percentage of lower status increases. In cases where there are large percentage of lower status, crime rate would increase with lower housing values. On the other hand, in cases where there are smaller percentage of lower status, lower housing values are

Figure 4. Interval Estimates of Correlation Coefficients



associated with lower crime rate, which make sense because housings with lower values might not be attractive to criminals in areas of less percentage of lower status.

TAX vs PTRAIO: The sample correlation is 0.4609. However, in areas where there is less percentage of lower status, the magnitude of correlation between property tax rate and pupil-teacher ratio is somewhat lower than the unconditional correlation estimate. The degree to which pupil-teacher ratio is associated with property-tax rate is lower in areas of less percentage of lower status compared to areas of large percentage of lower status. So, the unconditional correlation estimate seems to be more representative of the area

of large percentage of lower status than the area of lower percentage of lower status.

TAX vs MEDV: The sample correlation is -0.4685. However, Figure 4 shows a decreasing trend of correlations as the percentage of lower status increases. In cases where there are large percentages of lower status, the correlations between values of housings and property-tax rate tend to be highly negative, meaning that property-tax rate is negatively associated with housing values. On the other hand, in cases where there is less percentage of lower status, the correlation tends to be around zero, meaning that property-tax rate is not linearly associated with housing values. Thus, the unconditional sample correlation is more representative of the cases where there are large percentages of lower status.

PTRATIO vs MEDV: The sample correlation is -0.5078. Figure 4 shows that the conditional correlation is relatively stable across the percentage of lower status.

In conclusion the three correlations, corr(CRIM, MEDV), corr(TAX, PTRATIO), and corr(TAX, MEDV) are not well represented by the unconditional sample correlations. These findings are not available from the unconditional estimates of correlation coefficients.

6 Conclusion

Yin et al.'s (2010) approach is a good first step towards a better understanding of conditional covariance structure. However, their statistical estimation using NW kernel estimator has some limits.

First, Yin et al.'s estimator of covariance matrices is not unbiased. Its bias is of the order h^2 requiring researchers to use a small value of bandwidth in practice. However, it is not yet known what value of the bandwidth would be small enough or acceptable. There is a need for more studies to investigate this issue.

Second, Yin et al. considered a case of only one index variable, although there can be more than one. They argued that extension to multiple index variables is not practical due to the curse of dimensionality, although it is

theoretically straightforward. To overcome this issue, an aggregate multidimensional component can be considered. If the multiple index variables are correlated to some extent, it can be attempted to aggregate the index variables to a component that is multidimensional in nature. Then, the current approach can be applied. If the multiple index variables are not correlated, conditional covariances can be estimated separately for each index variable and they can be discussed accordingly.

Third, it is well known that the local linear model has better properties than the local constant model for kernel estimation: design adaptivity, correction of boundary effects, statistical efficiency in an asymptotic minmax sense (Fan, 1993). However, Yin et al. applied local constant model to estimate conditional covariance matrices to ensure positive definiteness of the estimated matrices. They posit that it is very challenging and not straightforward to develop an estimator of conditional covariance matrices following the local linear model without destroying the positive definiteness constraint. Then, we have to be prepared for the potential large bias at boundary points in the use of local constant model. To partially resolve this issue, Yin et al defined the index variable as "rank(LSTAT)/total sample size". As a result the index variable was forced to follow $U[0, 1]$. This procedure reduces bias because the second term of bias becomes 0 in the equation (18) (Yin et al., 2010). However, bias still remains as the first term is not zero in the equation (18). I ponder if it could be ensured that the second derivative of the covariance $\sigma_{j_1 j_2}(x)$ turns into a negligible value in future studies.

Fourth, Yin et al. does not provide a practical guide to evaluate the global convergence of $\hat{\sum}(x)$. Although they used Stein loss and quadratic loss, they did not give any value based on which practitioners can determine whether obtained values of Δ_1 and Δ_2 can be interpreted as a good convergence or not in estimation. To investigate this issue, there is a need for a large scale Monte-Carlo study.

References

[1] Andersen, T. G. and Lund, J. (1997). Estimating continuous time stochastic volatility models of the short term interest rate, *Journal of Econometrics*, **77**, 343-377.

[2] Benedetti, J. K. (1977). On the nonparametric estimation of regression functions. *Journal of Royal Statistical Society, Series B(Methodological).* **39**, 248-253.

[3] Bilmes, JA. (2000). Factorized sparse inverse covariance matrices. *IEEE International Conference on Acoustics, Speech, and Signal Processing.*

[4] Box, G. E. P. (1988). Signal-to-noise ratios, performance criteria, and transformations. *Technometrics*, **30**, 1-17.

[5] Bowman, A. W. and Azzalini, A (1997). *Applied smoothing techniques for data analysis: The Kernel approach with S-plus illustrations.* Oxford, NY: Oxford University Press.

[6] Bowman, A. W. and Azzalini, A (2018). *R package sm: nonparametri smoothing methods* (version 2.2-5.6) URL, http://www.stats.gla.ac.uk/ adrian/sm, http://azzalini.stat.unipd.it/Book sm.

[7] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association.* **87**, 829-836.

[8] Diggle P, Verbyla A. (1998). Nonparametric estimation of covariance structure in longitudinal data. *Biometrics* **54**, 401-415. [PubMed: 9629635]

[9] Drton M, Perlman M. (2004). Model selection for gaussian concentration graphs. *Biometrika* **91**, 591-602.

[10] Edwards, DM. (2000). Introduction to Graphical Modeling. Springer, New York.

[11] Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 998-1004.

[12] Fan, J. (1993). Local linear smoothers and their MINIMAX efficiencies. *Annals of Statistics*. **21**, 196-216.

[13] Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Annals of Statistics*. **20**, 2008-2036.

[14] Fan, J. and Gijbels, I. (1996). *Local polynomial modeling and its applications*. Capman and Hall, London.

[15] Gasser, T., Kneip, A., and Köhler, W. (1991). A flexible and fast method for automatic smoothing. *Journal of American Statistical Association*, **86**, 643-652.

[16] Gasser, T. and Müller, H-G. (1979). Kernel estimation of regression functions. In T. Gasser and M. Rosenblat Eds., *Smoothing techniques for curve estimation*, vol. 757 of Lecture notes in Mathematics, 23-68. Berlin: Springer.

[17] Györfi, L. Kohler, M., Krzyzak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. NY: Springer.

[18] Harrison, D. and Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, **5**, 81-102.

[19] Hollander, M., Wolfe, D. A., and Chicken, E. (2014). *Nonparametric Statistical Methods*, John Wiley: New York.

[20] Holst, U., Hössjer, O., Björklund, C., Ragnarson, P. and Edner, H. (1996). Locally weighted least squares kernel regression and statistical evaluation of LIDAR measurements. *Econometrics*, **7**, 401-416.

[21] Ledoit, O. and Wolf, M. (2017). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. Working paper series No. 122 (ISSN 1664-705X): Dept. of Economics, University of Zürich.

[22] Muirhead, R. J. (1982). Aspects of Multivariate Statistical Theory. John Wiley & Sons; New York.

[23] Nadaraya, E. (1964). On estimating regression. *Theory of Probability and its Applications* **9**, 141-142.

[24] Nadaraya, E. (1965). On non-parametric estimates of density functions and regression curves. *Theory of Probability and its Applications*, **10**, 186-190.

[25] Ruppert, D., Sheather, S. J., and Wand, M. D. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**, 1257-1270.

[26] Ruppert, D., Wand, M. P., Holst, U., and Hössjer, O. (1997). Local polynomial variance-function estimation. *Technometrics*, 39, 262-273.

[27] Scott, D. W. (2015) *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley: New York.

[28] Scott, D. W. (1983). Nonparametric Probability Density Estimation for Data Analysis in Several Dimensions. *Proceedings of the Twenty-Eighth Conference on the Design of Experiments in Army Research Development and Testing*. U.S. Army Research Office, pp.387-397.

[29] Scott, D. W. (1985). Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions. *Ann. Statist.* **13**. 1024-1040.

[30] Smith M, Kohn R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *J. Amer. Statist. Assoc.* **97**, 1141-1153.

[31] Watson, G. (1964). Smooth regression analysis. *Sankhyā : The Indian Journal of Statistics*, Series A. **29**, 359-372.

[32] Wood, S. N. (2017). Generalized additive models: An introduction with R. 2nd Ed. Boca Raton, FL: CRC Press.

[33] Yao, Q. and Tong, H. (1994). Quantifying the influence of initial values on nonlinear prediction. *Journal of Royal Statistics: Social*, B 56, 701-725.

[34] Yin, J., Geng, Z., Li, R., and Wang, H. (2010). Nonparametric Covariance Model. *Statistica Sinica*, **20**, 469-479.

[35] Yu, K. and Jones, M. C. (2004). Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association*, **99**, 139-144.

Appendix A. Proofs of Mean, Variance, and Bias for the kernel Estimator

The kernel estimator in $\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$ is the arithmetic mean of n independent and identically distributed random variables, $K_h = \frac{1}{h}K\left(\frac{x - X_i}{h}\right)$, which is a scaled outcome of $K\left(\frac{x - X_i}{h}\right)$ using h as the scale. Let's put $\frac{x - X_i}{h} = u$ for notational brevity. For further proofs we assume: $K(u) \geq 0,\ -\infty < u < \infty,\ K(-u) = K(u)$,

$$\begin{aligned}
&\int_{-\infty}^{\infty} K(u)du = 1,\\
&\int_{-\infty}^{\infty} uK(u)du = 0,\\
&\int_{-\infty}^{\infty} u^2K(u)du \equiv \sigma_K^2 > 0,\\
&\int_{-\infty}^{\infty} u^3K(u)du = 0,\\
&\int_{-\infty}^{\infty} u^4K(u)du \equiv \sigma_K^4.
\end{aligned} \tag{20}$$

Now, remember $K_h = \frac{1}{h}K\left(\frac{x - X_i}{h}\right)$. Then, $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h$.

$$E\hat{f}(x) = E\left[\frac{1}{n}\sum_{i=1}^{n} K_h\right] = \frac{1}{n}E\sum_{i=1}^{n} K_h = \frac{1}{n}n\left(EK_h\right) = EK_h$$

$$Var\ \hat{f}(x) = Var\left(\tfrac{1}{n}\sum_{i=1}^{n} K_h\right) = \tfrac{1}{n^2}n \cdot VarK_h = \tfrac{1}{n}VarK_h$$

$$
\begin{aligned}
EK_h &= \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x - X_i}{h}\right) f(X_i) dX_i, \quad \left(\frac{x - X_i}{h} = u, \frac{du}{dX_i} = -\frac{1}{h}, dX_i = -h du\right) \\
&\quad (X_i = \infty \Rightarrow u = -\infty; X_i = -\infty \Rightarrow u = \infty) \\
&= \int_{\infty}^{-\infty} \frac{1}{h} K(u) f(x - hu)(-h du) \\
&= \int_{-\infty}^{\infty} K(u) f(x - hu) du
\end{aligned} \tag{21}
$$

$x - hu = x - h(\frac{x - X_i}{h}) = x - (x - X_i) = X_i$
Then, $f(x - hu) = f(X_i)$ and $X_i - x = -hu$
Application of Taylor expansion to $f(X_i)$ at $X_i = x$ gives

$$
\begin{aligned}
f(X_i) &= \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!} (X_i - x)^n = \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!} (-hu)^n \\
&= f(x) - \frac{huf'(x)}{1} + \frac{(-hu)^2 f''(x)}{2!} + \cdots .
\end{aligned}
$$

Then equation (21) becomes

$$
\begin{aligned}
EK_h =& f(x) \int_{-\infty}^{\infty} K(u) du - hf'(x) \int_{-\infty}^{\infty} uK(u) du + \frac{1}{2} h^2 f''(x) \int_{-\infty}^{\infty} u^2 K(u) du \\
&- \frac{1}{3!} h^3 f^{(3)}(x) \int_{-\infty}^{\infty} u^3 K(u) du + \frac{1}{4!} h^4 f^{(4)}(x) \int_{-\infty}^{\infty} u^4 K(u) du + \cdots
\end{aligned} \tag{22}
$$

We apply assumptions (20) to equation (22). Then,

$$
\begin{aligned}
EK_h &= f(x) \cdot 1 - hf'(x) \cdot 0 + \frac{1}{2} h^2 f''(x) \cdot \sigma_K^2 - \frac{1}{3!} h^3 f^{(3)}(x) \cdot 0 + \frac{1}{4!} h^4 f^{(4)}(x) \sigma_K^4 + \cdots \\
&= f(x) + O(h^2)
\end{aligned} \tag{23}
$$

Then

$$E\hat{f}(x) = EK_h = f(x) + O(h^2). \tag{24}$$

Conditions required to guarantee the convergence of $f(X_i) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!}(-hu)^n$ will be discussed.

Suppose that we have $\sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!}(-h)^n \cdot u^n = \sum a_n u^n$.

Since this is a power series, it converges if $\lim_{n\to\infty} |\frac{a_{n+1}u^{n+1}}{a_n u^n}| < 1$, given $h \neq 0$ and $u \neq 0$.

$$\begin{aligned}\lim_{n\to\infty} |\frac{a_{n+1}u^{n+1}}{a_n u^n}| &= \lim_{n\to\infty} |\frac{f^{(n+1)}(x)(-h)^{n+1}u^{n+1}n!}{(n+1)!f^{(n)}(x)(-h)^n u^n}| \\ &= \lim_{n\to\infty} |\frac{f^{(n+1)}(x)(-h)u}{f^{(n)}(x)(n+1)}| < 1, f^{(n)}(x) \neq 0.\end{aligned}$$

Then, $|u||\frac{f^{(n+1)}(x)}{f^{(n)}(x)}| \lim_{n\to\infty} \frac{1}{n+1} = 0 < 1$.

So, $f^{(n)}(x) \neq 0$ is the first condition to ensure the convergence of $f(X_i) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!}(-hu)^n$.

The second condition for the convergence will be derived.

$$\begin{aligned}f(X_i) &= \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!}(-hu)^n \\ &= f(x) - \frac{huf'(x)}{1} + \frac{(-hu)^2 f''(x)}{2!} + \cdots \frac{(-hu)^n f^{(n)}(x)}{n!} + R_n(X_i),\end{aligned}$$

where $R_n(X_i) = \frac{(-hu)^{n+1} f^{n+1}(x)}{(n+1)!}$.

Suppose we have that $|f^{n+1}(c)| \leq M$ for all c between x and X_i, where M is a positive constant. Then, $|R_n(X_i)| = \frac{|(-hu)|^{n+1}|f^{n+1}(x)|}{(n+1)!} \leq M\frac{|(-hu)|^{n+1}}{(n+1)!}$.

Letting $n \to \infty$, $\frac{|(-hu)|^{n+1}}{(n+1)!} \to 0$. Thus, $|R_n(X_i)| \to 0$ by the comparison theorem. So, $\lim_{n\to\infty} R_n(X_i) = 0$. And the power series of $f(X_i)$ converges to $f(X_i)$ for every X_i. Thus, for every n and a positive constant M, $|f^{(n+1)}(c)| \leq M$ for all c between x and X_i is the second condition for the power series to converge to $f(X_i)$.

$$VarK_h = EK_h^2 - (EK_h)^2 \tag{25}$$

$$\begin{aligned}
EK_h^2 &= \int_{-\infty}^{\infty} K_h^2 f(X_i) dX_i = \int_{-\infty}^{\infty} \frac{1}{h^2} K(\frac{x - X_i}{h})^2 f(X_i) dX_i \\
&\left(\frac{x - X_i}{h} = u, \frac{du}{dX_i} = -\frac{1}{h} \right) \\
&= \frac{1}{h^2} \int_{\infty}^{-\infty} K(u)^2 f(x - hu)(-h\, du) = \frac{1}{h} \int_{-\infty}^{\infty} K(u)^2 f(x - hu) du \\
&[Taylor\, expansion\, of (x - hu)\ with\ f^{(n)}(x) \neq 0 : f(x) - \frac{huf'(x)}{1} + \frac{(-hu)^2 f''(x)}{2} + \cdots] \\
&= \frac{1}{h}[f(x) \int_{-\infty}^{\infty} K(u)^2 du - hf'(x) \int_{-\infty}^{\infty} uK(u)^2 du + \frac{1}{2} h^2 f''(x) \int_{-\infty}^{\infty} u^2 K(u)^2 du \\
&- \frac{1}{3!} h^3 f^{(3)}(x) \int_{\infty}^{\infty} u^3 K(u)^2 du + \frac{1}{4!} h^4 f^{(4)}(x) \int_{\infty}^{\infty} u^4 K(u)^2 du + \cdots] \\
&= \frac{1}{h}[f(x)R(K) + O(h)] \approx \frac{1}{h} f(x) R(K), \text{ where } R(K) = \int_{-\infty}^{\infty} K(u)^2 du.
\end{aligned} \tag{26}$$

Substituting equation (23) and equation (26) into equation (25),

$$VarK_h = EK_h^2 - (EK_h)^2 = \frac{f(x)R(K)}{h} - [f(x) + O(h^2)]^2$$

$$
\begin{aligned}
Var\,\hat{f}(x) &= \frac{1}{n} VarK_h = \frac{f(x)\cdot R(K)}{n\cdot h} - \frac{[f(x)+O(h^2)]^2}{n} \\
&= \frac{f(x)R(K)}{n\cdot h} - \frac{1}{n}\left(f(x)^2 + 2f(x)\cdot O(h^2) + (O(h^2))^2\right) \\
&= \frac{f(x)R(K)}{n\cdot h} - \frac{1}{n}(f(x)^2 + O(h^2)) \\
&= \frac{f(x)R(K)}{n\cdot h} - \frac{f(x)^2}{n} + O(\frac{h}{n})
\end{aligned}
\tag{27}
$$

From equations (24) and (27), $E\hat{f}(x)$ and $Var\,\hat{f}(x)$ are given.
Using equation (23) and equation (27),

$$
\begin{aligned}
Bias\,\hat{f}(x) &= E\,\hat{f}(x) - f(x) = EK_h - f(x) \\
&= f(x) + \frac{1}{2}h^2 f''(x)\sigma_K^2 + O(h^4) - f(x) \\
&= \frac{1}{2}h^2 f''(x)\sigma_K^2 + O(h^4)
\end{aligned}
$$

$$
(Bias\,\hat{f}(x))^2 = \frac{1}{4}h^4[f''(x)]^2\sigma_K^4 + O(h^6)
$$

$$
MSE = Bias^2 + Var = \frac{1}{4}h^4[f''(x)]^2\sigma_K^4 + O(h^6) + \frac{f(x)R(K)}{nh} - \frac{f(x)^2}{n} + O(\frac{h}{n}).
$$

Appendix B. Procedures of Developing Sample Data for Simulation

In this appendix, n stands for sample size, ne for number of evaluation points, p for number of study variables, and $nrep$ for number of replications.
(1) The random index variable is denoted by U. We develop n values of $u \sim U(0,1)$ for one sample. A matrix $uv(n \times nrep)$ is obtained containing a vector of n values of u for each replication.
(2) Pick one value of u from the column of the uv matrix. Using the value, we develop a mean vector and a covariance matrix that will be used to generate a sample observation in the next step.
(3) Generate one observation from a multivariate population with the mean vector and the covariance matrix developed in the step (2).
(4) Repeat step (3) until n observations are collected to make a sample dataset of five study variables for a replication: An array data.x ($n \times p \times \ nrep$) is obtained.
(5) Randomly select $ne = 50$ points on the continuum of [0, 1]. Then a matrix $ue(ne \times nrep)$ is obtained containing a vector of ne values for each replication.
(6) Pick one value from the column of the ue matrix. Using the value, we develop a mean vector and a covariance matrix that will be used to generate a sample observation in the next step.
(7) Generate one observation from a multivariate population with the mean vector and the covariance matrix developed in the step (6).
(8) Repeat step (7) until ne observations are collected to make a sample dataset for each of evaluation points in a replication: An array data.xe($ne \times p \times nrep$) is obtained.