A data-driven method for estimating conditional densities

By JIANQING FAN

Department of Statistics, Chinese University of Hong Kong, Shatin, Hong Kong.

and TSZ HO YIM

Department of Statistics, Chinese University of Hong Kong, Shatin, Hong Kong.

Summary

The conditional density function is very useful for forecasting and statistical inferences, particularly in financial economics. Yet, its bandwidth selection has not yet been systematically studied. In this article, we extend the idea of cross-validation (CV) for choosing the smoothing parameter of the "double-kernel" local linear regression for estimating a conditional density. Our selection rule optimizes the estimated conditional density function by minimizing the integrated square error (ISE). We also discuss three other bandwidth selection rules. The first is an ad-hoc method used by Fan, Yao and Tong (FYT, 1996). The second rule, as suggested by Hall, Wolff and Yao (HWY, 1999), employs the idea of bootstrap for the bandwidth selection in the estimation of conditional distribution functions. We modify the HWY approach to suit the bandwidth selection for the conditional density function. The last is the rule of thumb approach proposed by Hyndman and Yao (2002). The performance of the newly proposed CV approach is compared with these three methods by simulation studies, and our method performs outstandingly. The method is also illustrated by application to two sets of time series.

Some key words: Bandwidth selection; Bootstrap; Conditional density function; Cross validation; Local-linear method.

1 Introduction

Conditional density provides the most informative summary of the relationship between independent and dependent variables. Its center indicates the regression function and its dispersion shows the likely size of prediction errors. When the conditional density appears asymmetric or multimodal, the mean regression function and its associated standard deviation are not adequate to summarize the conditional distribution. In contrast, the conditional density function provides an informative summary; see, for example, the monographs by Silverman (1986) and Wand and Jones (1995).

In stationary time series with Markovian structures, conditional density characterizes the probabilistic aspect of the time series. It determines, except for the initial distribution, the likelihood

function of an observed time series and facilitates statistical prediction (Tong, 1990 and Fan and Yao, 2003). It is very useful for studying nonlinear phenomena, such as the symmetry, multimodality structure, and sensitivity measures of a time series (Chan and Tong, 2001).

The conditional density function plays a pivotal role in financial economics. It is directly related to the pricing formula of financial derivatives and inferences of parameters in financial models (Aït-Sahalia, 1999). Consider the continuous time model in which an economic variable, X_t, satisfies the following stochastic difference equation:

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t, \tag{1.1}$$

where W_t is a standard Brownian motion and μ and σ^2 are drift and diffusion functions. This continuous time model (1.1) is a basic stochastic dynamic model and is widely used in finance and economics. It includes many well known single-factor models such as those of Black and Scholes (1973), Vasicek (1977), Cox, Ingersoll and Ross (1980, 1985), and Chan, Karolyi, Longstaff and Sanders (1992) for modeling stock prices or interest rates. These models, except the last, admit close forms of the conditional probability density function, which allows us to explicitly evaluate the price of financial derivatives based on X_t. For other cases, approximations of conditional densities are needed (see Aït-Sahalia, 1999).

The nonparametric estimation of the drift and diffusion functions in (1.1) based on discretely observed data (with a fixed sampling frequency) is an interesting and challenging problem. Interest in the problem has recently surged in the literature on financial econometrics and statistics. For an overview of and references for nonparametric techniques in financial economics, see the work of Fan (2003). The transition density allows one to estimate the unknown functions from model (1.1). In fact, it determines the drift function $\mu(\cdot)$ and diffusion function $\sigma(\cdot)$ in model (1.1). See, for example, the work of Hansen, Scheinkman and Touzi (1998) for a spectral approach to such a determination. Thus, a nonparametric estimate of the transition density based on discretely observed time series data allows us to estimate and make inferences about the drift and diffusion functions, including checking the validality of the well-known financial models that are mentioned above.

A vast variety of papers use the estimators of conditional densities as building blocks. Such papers include those of Robinson (1991), Tjϕstheim (1994), Polonik and Yao (2000), among others. However, in all of those papers, the conditional density function was indirectly estimated. Hyndman, Bashtannyk and Grundwald (1996) studied the kernel kerstimator of conditional density estimator and its bias corrected version. Fan, Yao and Tong (1996) developed a direct estimation method via an innovative "double-kernel" local linear approach that connected the estimation of the conditional density function to the nonparametric regression problem. There are many advantages of using local linear regression, such as the lack of boundary modifications, high minimax efficiency, ease of implementation, etc. More details of this method can be found in the book by Fan and Yao (2003).

Despite its importance in various applications, the problem of automatically estimating the

conditional density function has not systematically studied. After the pioneering work of Rosenblatt (1969), systematic studies of this subject were infrequent until it was revisited recently. In particular, to our knowledge, no consistent data-driven selection procedure has been proposed for choosing smoothing parameters.

The smoothing parameter or the bandwidth plays an important role in the performance of local linear regression estimators as it controls the model complexity. A too small bandwidth results in undersmoothing and tends to yield a wiggly curve. A too large bandwidth results in oversmoothing so that important features have been smoothed away. Data-driven methods attempt to balance these two problems by selecting an appropriate bandwidth. For the estimation of conditional density function, two smoothing parameters, namely h_1 and h_2, are required, where h_1 controls the smoothness between conditional densities in the x-direction and h_2 controls the complexity of each conditional density in the y-direction.

Fan, Yao and Tong (1996), hereafter abbreviated as FYT, provided estimators for a conditional density function based on a "double-kernel" local linear approach using an ad-hoc method for selecting the smoothing parameters. Hall, Wolff and Yao (1999), hereafter abbreviated as HWY, suggested, for a related topic, a bandwidth selection method for estimating a conditional distribution function. A bootstrap approach was also introduced for the bandwidth selection. Bashtannyk and Hyndman (2001) and Hyndman and Yao (2002) proposed several useful rules of thumbs for selection bandwidths. All of the above methods for bandwidth selection are simple and ad hoc, and cannot be consistent.

The goal of this paper is to develop a consistent data-driven bandwidth selection rule for estimating conditional density functions. This bandwidth selection rule is based on the cross-validation (CV) method. Although this paper is motivated by the problem of time series data, we present our method in a more general situation so that the time series model is a particular case.

The organization of this paper is as follows. In Section 2, we outline the key ideas of the estimation of conditional density using the double-kernel local linear regression technique. We then introduce three bandwidth selection rules for estimating the conditional density function in Section 3: the FYT approach, the bootstrap (HWY) approach, and the newly proposed CV method. Simulation studies are provided in Section 4. The applications of proposed methods to two sets of real data are given in Section 5. All of the technical proofs are given in the Appendix.

2 Estimation of conditional density

We assume that the data are available in the form of a strictly stationary process (X_i, Y_i) with the same marginal distribution as (X, Y). Naturally, this includes the case in which the data (X_i, Y_i) are independent and identically distributed. Let $g(y|x)$ be the conditional density of Y given $X = x$, evaluated at $Y = y$. This conditional density can be estimated via the "double-kernel" local linear method of Fan, Yao and Tong (1996).

Estimating the conditional density can be regarded as a nonparametric regression problem. To make this connection, FYT observed that as $h_2 \to 0$,

$$E\{K_{h_2}(Y-y)|X=x\} \approx g(y|x), \tag{2.1}$$

where K is a nonnegative density function and $K_h(y) = K(y/h)/h$. The left hand side of (2.1) is the regression function of the random variable $K_{h_2}(Y-y)$ given $X=x$. By Taylor's expansion about x, we have

$$\begin{aligned} E\{K_{h_2}(Y-y)|X=z\} &\approx g(y|z) \\ &\approx g(y|x) + g'(y|x)(z-x) \\ &\equiv \alpha + \beta(z-x), \end{aligned}$$

where $g'(y|x)$ is the partial derivative of $g(y|x)$ with respect to x.

For each given x and y, the principle of the local linear regression suggests the minimization of

$$\sum_{i=1}^{n} \{K_{h_2}(Y_i - y) - \alpha - \beta(X_i - x)\}^2 W_{h_1}(X_i - x), \tag{2.2}$$

with respect to the local parameters α and β, where W is a nonnegative density function. The resulting estimate of the conditional density is simply $\hat{\alpha}$.

It is more convenient to work with matrix notation. Let $\mathcal{X}$ be the design matrix of the local least-squares problem (2.2) and

$$\mathcal{W} = \mathrm{diag}(W_{h_1}(X_1 - x), \ldots, W_{h_1}(X_n - x)),$$

$$\mathcal{Y} = (K_{h_2}(Y_1 - y), \ldots, K_{h_2}(Y_n - y))^T.$$

We define

$$S_n(x) = n^{-1}\mathcal{X}^T\mathcal{W}\mathcal{X} \quad \text{and} \quad T_n(x,y) = n^{-1}\mathcal{X}^T\mathcal{W}\mathcal{Y}.$$

Then, by simple algebra, the estimated conditional density can be expressed as

$$\hat{g}_h(y|x) = e_1^T S_n^{-1}(x) T_n(x,y), \tag{2.3}$$

where $e_1^T = (1,0)$ and $h = (h_1, h_2)^T$.

It is also instructive to express the estimated conditional density in the form of the equivalent kernel. Let

$$W_n(z;x) = e_1^T S_n^{-1}(x) \begin{pmatrix} 1 \\ zh_1 \end{pmatrix} W(z), \tag{2.4}$$

be the equivalent kernel. Then, the estimator (2.3) can be written as

$$\hat{g}_h(y|x) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} W_n\left(\frac{X_i - x}{h_1}; x\right) K\left(\frac{Y_i - y}{h_2}\right). \tag{2.5}$$

The effective kernel (2.4) can be explicitly expressed as (see Fan and Yao, 2003)

$$W_n(z;x) = W(z)\frac{s_{n,2}(x) - zh_1 s_{n,1}(x)}{s_{n,0}(x)s_{n,2}(x) - s_{n,1}(x)^2},$$

where

$$s_{n,j}(x) = n^{-1}\sum_{i=1}^{n}(X_i - x)^j W_{h_1}(X_i - x), \qquad j = 0, 1, 2.$$

3 Bandwidth selection

3.1 *A rule of thumb*

Fan, Yao and Tong (1996) proposed an ad-hoc rule of thumb for selecting the smoothing parameters. For simplicity, in the density estimation setting, the normal referencing rule (Silverman, 1986, p. 45) selects the bandwidth

$$\hat{h}_2 \quad = \quad \left[\frac{8\pi^{1/2}\int K^2(x)dx}{3(\int x^2 K(x)dx)^2}\right]^{1/5} s_y n^{-1/5}, \tag{3.1}$$

where s_y is the sample standard deviation of Y.

For given bandwidth h_2 and y, (2.2) is a standard nonparametric problem of regressing $K_{h_2}(Y_i - y)$ on X_i. There are many data-driven methods for selecting the bandwidth h_1. These include cross-validation (Stone, 1974), the residual-square criterion (Fan and Gijbels, 1995), the pre-asymptotic substitution method (Fan and Gijbels, 1995), the plug-in method (Ruppert, Sheather and Wand, 1995) and the empirical bias method (Ruppert, 1997), among others. For ease of reference, we will call this method for selecting the smoothing parameters h_1 and h_2 as the FYT approach. This is only a simple rule of thumb, and is not expected to work in all situations. In our numerical implementations, we use the plug-in method of Ruppert, Sheather and Wand (1995) to choose h_1. As shown below, the FYT approach tends to oversmooth the conditional density in the y-direction.

Bashtannyk and Hyndman (2001) proposed a rule of thumb estimator based on the assymption that the conditonal density is normal with linear regression and linear conditional standard deviation estimator and that the marginal density is either uniform or truncated normal. They further improved the procedure by combining the idea with that of FYT. Hyndman and Yao (2002) introduced new conditional density estimators based on local polynomial fit with nonnegativity constraints and derived also several rules of thumb bandwidth selectors using a similar idea. When W and K are guassian kernel, they suggested using

$$h_1 = 0.935(\nu\sigma^5/n|d_1|^5)^{1/6} \quad \text{and} \quad h_2 = |d_1|h_1$$

by assuming that the conditional density is $N(d_0 + d_1 x, \sigma^2)$ and the marginal density of X is $N(\mu, \nu^2)$. For simplicity, we will refer to this approach as the HY method.

3.2 *Bootstrap bandwidth selection*

Instead of using the FYT approach, Hall, Wolff and Yao (1999) suggested a bootstrap approach for selecting the smoothing parameters in the context of estimating the conditional distribution functions. Their idea can be adapted here to select h_1 and h_2 for estimating the conditional density function. We now introduce the idea. First, fit a simple parametric model

$$Y_i = a_0 + a_1 X_i + \cdots + a_k X_i^k + \sigma\epsilon_i, \quad \epsilon_i \sim N(0,1)$$

where $a_0, \ldots, a_k$, and σ are estimated from the data and k is determined by the Akaike information criterion (AIC). A parametric estimator $\check{g}(y|x)$ is then formed based on the selected parametric model. For $i = 1, \ldots, n$, by the Monte Carlo simulation, generate $\epsilon_i^* \sim N(0,1)$ and compute

$$Y_i^* = \hat{a}_0 + \hat{a}_1 X_i + \cdots + \hat{a}_{\hat{k}} X_i^{\hat{k}} + \hat{\sigma}\epsilon_i^*. \tag{3.2}$$

Hence, we obtain a bootstrap sample of $\{Y_1^*, \ldots, Y_n^*\}$ and a bootstrap estimate, $\hat{g}_h^*(y|x)$, that is derived from (2.5) with $\{(X_i, Y_i)\}$ replaced by $\{(X_i, Y_i^*)\}$. Let

$$M(h; x, y) = E[|\hat{g}_h^*(y|x) - \check{g}(y|x)||\{(X_i, Y_i)\}]$$

be the bootstrap estimator of the absolute deviation error of $\check{g}(y|x)$. The expectation is taken with respect to the bootstrap sample and can be computed by simulation. Finally, we choose h to minimize

$$M(h) = \frac{1}{n}\sum_{i=1}^{n} M(h; X_i, Y_i) I(X_i \in [a, b]),$$

where $[a, b]$ is an interval at which we want to estimate the conditional density. Again, this method is expected to work well for polynomial regression models and cannot be consistent for other models. For ease of reference, we call this method the HWY approach.

3.3 *A cross-validation method*

The FYT and HWY approaches are simple and ad hoc methods. They are not intended to optimize the estimated conditional densities. In fact, the smoothing parameters h_1 and h_2 in the FYT approach are selected separately. The bootstrap method provides good approximation when the true model is normal and the regression function is polynomial. However, in real situations, the true model can be asymmetric or heavy-tailed. For these situations, the bootstrap method fails to select the optimal bandwidths. We here extend a cross-validation (CV) idea of Rudemo (1982) and Bowman (1984) for estimating conditional density.

Let $f(x)$ be the marginal density of $\{X_i\}$ and $[a,b]$ be an interval at which we wish to estimate the conditional density. Define the Integrated Square Error (ISE) as

$$\begin{aligned} & \int \{\hat{g}_h(y|x) - g(y|x)\}^2 f(x) I(x \in [a,b]) dxdy \\ = & \int \hat{g}_h(y|x)^2 f(x) I(x \in [a,b]) dxdy - 2 \int \hat{g}_h(y|x) g(y|x) f(x) I(x \in [a,b]) dxdy \\ & + \int g(y|x)^2 f(x) I(x \in [a,b]) dxdy. \end{aligned} \tag{3.3}$$

Notice that the last term does not depend on h. To minimize the ISE with respect to h, we can ignore the third term.

A reasonable estimate in (3.3) is

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \in [a,b]) \int \hat{g}_{h,-i}(y|X_i)^2 dy - \frac{2}{n} \sum_{i=1}^{n} \hat{g}_{h,-i}(Y_i|X_i) I(X_i \in [a,b]), \tag{3.4}$$

where $\hat{g}_{h,-i}(y|x)$ is the estimator of (2.5) based on the sample $\{(X_j, Y_j), j \neq i\}$. Note that the first integral in (3.4) can be explicitly calculated with (2.5).

Estimating conditional density is much more involved than density estimation setting and the ISE cannot be estimated without bias. In fact, the bandwidths h_1 and h_2 play very different roles in the smoothing. It is not clear whether the proposed CV method is reasonable. To appreciate how much the bias the $CV(h)$ involves, we would like to compute the expected value of $CV(h)$. However, this is not viable in the regression setting due to the random denominator. Instead, we can compute the conditional expectation. However, for the times series applications, the design points $\{X_t, t = 1, \cdots, T-1\}$ involve nearly whole series. Hence, this approach is not applicable. The device of asymptotic normality is frequently used to avoid this kind of difficulty. See, for example, Chapter 6 of Fan and Yao (2003). While this can be done in the current context, it would involve substantial technicality. To mitigate the technicality and highlight the key insight, we appeal to the independent random sample setting.

Let $\{(X_i, Y_i), i = 1, \cdots, n\}$ be a random sample from a population with conditional density $g(y|x)$ and design density $f(x)$. For any random variable, let $E_X(Z)$ be the conditional expectation of Z given $X_1, \cdots, X_n$, namely

$$E_X(Z) = E(Z|X_1, \cdots, X_n).$$

The following result will be proved in the Appendix.

THEOREM 1. *Assume that the kernels K and W are bounded with bounded supports and vanishing first moments. If $f(\cdot)$ is continuous and positive in an open interval containing $[a,b]$ and $p(\cdot|x)$ is bounded for x in an open interval containing $[a,b]$, then*

$$\begin{aligned} & E_X \frac{1}{n} \sum_{i=1}^{n} \hat{g}_{h,-i}(Y_i|X_i) I(X_i \in [a,b]) = \\ & \quad E_X \int \hat{g}_h(y|x) g(y|x) f(x) I(x \in [a,b]) dxdy + O_P\left(\frac{1}{nh_1}\right). \end{aligned} \tag{3.5}$$

Furthermore, under the additional condition that $f'(\cdot)$ exists and is continuous in an open interval containing $[a, b]$, then

$$E_X \frac{1}{n} \sum_{i=1}^{n} I(X_i \in [a,b]) \int \hat{g}_{h,-i}(y|X_i)^2 dy =$$
$$E_X \int \hat{g}_h(y|x)^2 f(x) I(x \in [a,b]) dx dy + O_P\left(\frac{1}{nh_1}\right), \tag{3.6}$$

The biases in (3.5) and (3.6) are negligible to the first order since the variance of $\hat{g}_h(y|x)$ is of order $O_P\left(\frac{1}{nh_1h_2}\right)$ (See Fan, Yao and Tong, 1996).

4 Simulation studies

We consider simulation studies to evaluate and compare bandwidth selection methods for estimating the conditional density that were described in Section 3. These selection rules are the cross-validation method (CV), the bootstrap approach (HWY), the FYT and HY methods. For each simulation, the performance of the selection rule is evaluated by the root-mean square error (RMSE)

$$\text{RMSE} = \frac{\sqrt{\sum_i (\hat{g}_h(y_i|x_i) - g(y_i|x_i))^2 I(x_i \in [a,b])}}{\sum_i I(x_i \in [a,b])}$$

where (x_i, y_i) are grid points that are evenly distributed across certain regions of interest and [a, b] is an interval in the x-direction on which we wish to estimate the conditional density. We let K and W be the Gaussian kernel throughout this section.

Example 1. (Location Model) We consider a simple quadratic model

$$Y_i = 0.23 X_i (16 - X_i) + 0.4\epsilon_i, \quad i \geq 1.$$

The noise term is simulated from the following situations:

a) $\{\epsilon_i\}$ are independent standard normal random variables;

b) $\{\epsilon_i\}$ are a random sample from t_2-distribution; and

c) $\{\epsilon_i\}$ are independent and follow t_4-distribution.

For cases (a)-(c), X_i are independent uniform random variables on $[0, 16]$. We also consider the following time series setting:

d) $X_i = Z_{i-1}$, $Y_i = Z_i$ with some initial values Z_0. The noise ϵ_i are independent random variables with the same distribution as the random variable η, which is equal to the sum of 48 independent random variables each uniformly distributed on $[-0.25, 0.25]$. According to the central limit theorem, ϵ_i can be treated as nearly a standard normal variable. However, it has a bounded support $[-12, 12]$. Note that the bounded support of ϵ_i is necessary for the stationarity of the time series (cf. Chan and Tong 1994). The conditional density of this model was studied by Fan, Yao and Tong (1996).

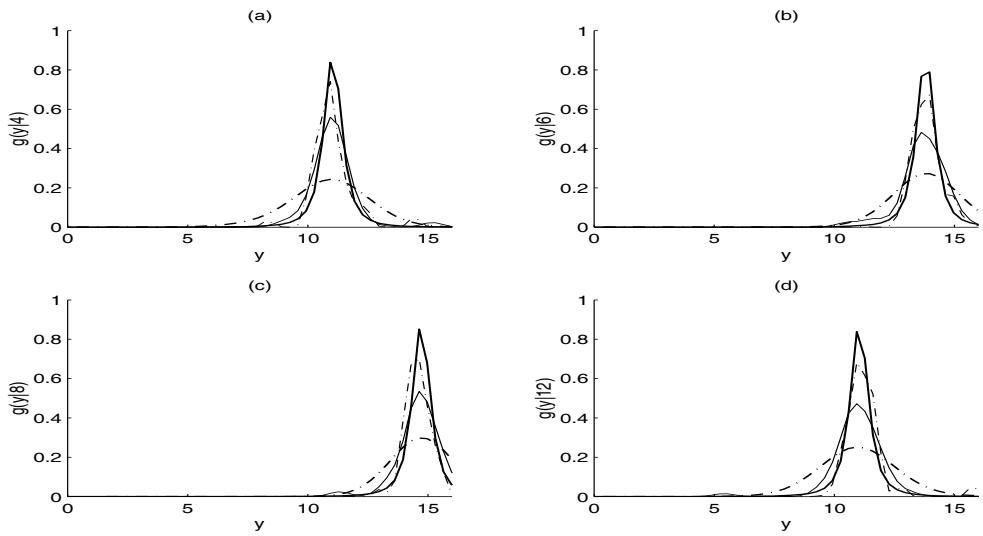

Figure 1: Estimated conditional densities for Example 1(b). Estimated conditional densities for (a) $x = 4$, (b) $x = 6$, (c) $x = 8$, and (d) $x = 12$ using the CV (thin dashdot curve), HWY (thin solid curve), and FYT (thick dashdot curve) approaches along with the true densities (thick solid curve).

For each of the 100 samples of size $n = 1000$, we calculate the RMSEs with different bandwidth selection rules. We estimate $g(y|x)$ on a 51×51 regular grid on the sample space. We take $a = 2$ and $b = 14$ for cases (a)-(c), and $a = 4$ and $b = 14$ for case (d).

Table 1: Summary of the RMSE for Example 1.

	case (a)	case (b)	case (c)	case (d)
CV	$1.0899^a(0.0601^b)$	0.7641(0.1497)	1.0143(0.0683)	1.0903(0.0673)
	$1.0882^c(0.0616^d)$	0.7842(0.1277)	1.0141(0.0712)	1.0888(0.0679)
HWY	1.0651(0.0516)	1.0191(0.1251)	1.0200(0.0651)	1.0803(0.0690)
	1.0631(0.0515)	1.0276(0.1095)	1.0194(0.0633)	1.0774(0.0665)
FYT	2.8121(0.0254)	1.6773(0.2908)	2.3902(0.0810)	2.7301(0.0515)
	2.8132(0.0254)	1.7441(0.2421)	2.4100(0.0690)	2.7300(0.0451)
HY	3.2158(0.0247)	1.8855(0.3392)	2.7433(0.0912)	3.0160(0.0508)
	3.2172(0.0234)	1.9577(0.2954)	2.7644(0.0795)	3.0129(0.0521)

NOTE: Mean^a(SD^b) and median^c(robust SD$(=\frac{\text{IQR}}{1.35})^d$) of RMSE$(\times 10^{-3})$ for each method.

We summarize the results numerically in Table 1. The means, standard derivations, medians, and robust standard derivations of the RMSEs of each of the three bandwidth selection rules are given. In general, the HY and FYT approaches produce larger RMSEs than those of CV and the HWY method. The performance of both CV and the HWY method are comparable. Cases (a) and (d) are ideal for the HWY approach because this is the model in which the bootstrap

sample was generated. Nevertheless, the proposed CV approach works comparably to the HWY method, which is a parametric approach for this model. However, the performance of the HWY approach deteriorates when the tails of the noise ϵ_i become heavier. See case (b) and Figure 1, which presents a typical estimated conditional density with median performance. In fact, for case (b) the CV method outperforms the other two approaches substantially.

Example 2. (Location Model) Let us consider the cosine model

$$Y_i = 20\cos\left(\frac{\pi X_i}{10}\right) + \epsilon_i, \quad i \geq 1$$

where

a) X_i are independent uniform random variables on $[-20, 20]$;

b) $X_i = Z_{i-1}$, $Y_i = Z_i$ with some initial values Z_0.

For both cases (a) and (b), the noise ϵ_i are independent standard normal random variables. Case (b) was studied by Fan, Yao and Tong (1996).

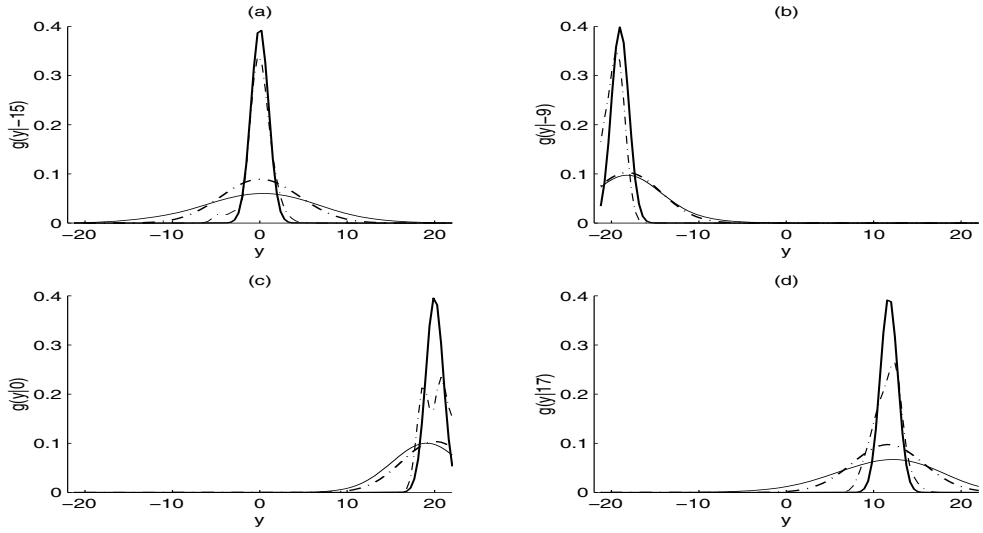

Figure 2: Estimated densities for Example 2(b) at (a) $x = -15$, (b) $x = -9$, (c) $x = 0$, and (d) $x = 17$ using the CV (thin dashdot curve), HWY (thin solid curve), and the FYT (thick dashdot curve) approaches along with the true densities (thick solid curve).

The sample size n is 1000 and the number of the replications in the simulation is 100. We took $a = -18$ and $b = 18$ for case (a) and $a = -17$ and $b = 20$ for case (b). Table 2 summarizes the results numerically and Figure 2 depicts the results of case (b) graphically. The plots show the true and estimated conditional densities at $x = -15$, $x = -9$, $x = 0$ and $x = 17$.

The performance of the CV approach is much better than that of the other three approaches both in the numerical and graphical summary. In this example, the HWY approach gives estimates that are oversmooth in the y-direction. The FYT method somewhat outperforms the HWY approach.

Table 2: Summary of the RMSE for Example 2.

	case (a)				case (b)			
	CV	HWY	FYT	HY	CV	HWY	FYT	HY
Mean	2.7404	8.0545	7.2869	9.3314	3.1282	8.2218	7.6028	8.7231
SD	0.1262	0.1237	0.1060	0.1544	0.1464	0.1201	0.1085	0.1142
Median	2.7182	8.0543	7.2953	9.3279	3.1238	8.2259	7.6091	8.7229
Robust SD	0.1261	0.1156	0.1071	0.1604	0.1721	0.1156	0.1164	0.1228

NOTE: Mean, SD, median, and robust SD(=IQR/1.35) of RMSE($\times 10^{-3}$) for each method.

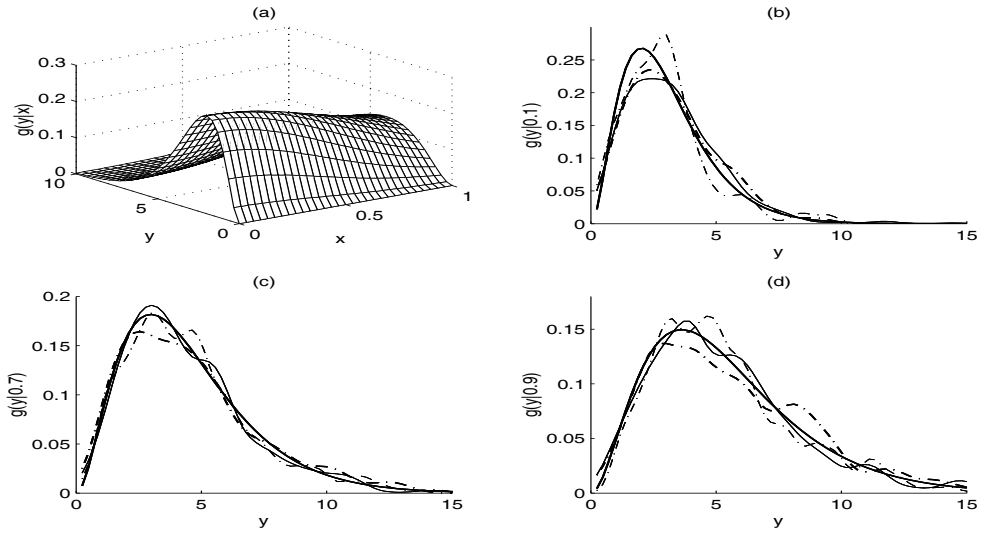

Figure 3: Graphical summary for Example 3. (a) The true conditional density function. (b)-(d) Estimated densities for (b) $x = 0.1$, (c) $x = 0.7$, (d) $x = 0.9$ using CV (thin dashdot curve), HWY (thin solid curve), FYT (thick dashdot curve) approaches along with the true densities (thick solid curve).

The model used in this example deviates from the bootstrap model and this is the main reason for the poor performance of the HWY method. This shows that the HWY method is not consistent.

Table 3: Numerical summary for Example 3.

	mean	SD	median	robust SD
CV	1.7885	0.3207	1.7014	0.3937
HWY	2.0113	0.3286	2.0410	0.3450
FYT	2.0380	0.4278	2.0302	0.4071
HY	1.7956	0.3081	1.7574	0.3854

Example 3. (Scale Model) Here we consider the following conditional density function:

$$Y_i|X_i \sim Gamma(3, X_i^2 + 1), \quad i \geq 1$$

where X_i are independent standard uniform random variables.

A sample of 1000 was generated from the above model and we repeated the simulation 100 times. We took $a = 0.1$ and $b = 0.9$ and estimated $g(y|x)$ on regular grid points with steps 0.016 and 0.5949 in the x- and y-directions. Table 3 summarizes the results numerically and Figure 3 gives the graphical results. Plot (a) shows the true conditional density and the remaining plots show the true and estimated densities at $x = 0.1$, $x = 0.7$, $x = 0.9$. The values of the x's are chosen so that the difference in variance between consecutive x's is roughly the same. Both the numerical and graphical results show that the performance of the above four bandwidth rules is comparable, though the CV and HY methods have some advantages.

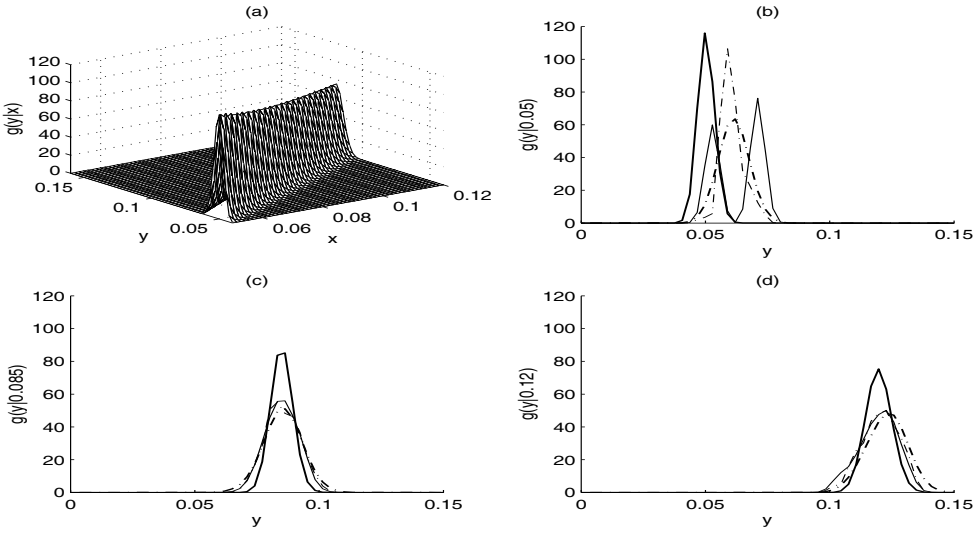

Figure 4: Two-step transition densities for the CIR model. (a) The true conditional density function. (b)-(d) Estimated densities for (b) $x = 0.07$, (c) $x = 0.1$, and (d) $x = 0.13$ using the CV (thin dashdot curve), HWY (thin solid curve), and FYT (thick dashdot curve) approaches, and compared with the true densities (thick solid curve).

Example 4. (CIR Model) We consider the well known CIR model for modeling the interest rate dynamic structure:

$$dX_t = \kappa(\theta - X_t)dt + \sigma\sqrt{X_t}dW_t, \quad t \geq t_0. \tag{4.1}$$

The CIR model is an example of model (1.1). The interest rate X_t moves around a central location or long-run equilibrium level θ. When $X_t > \theta$, a negative drift pulls it down, and when $X_t < \theta$, a positive force drives it up. The parameter κ determines its speed. If $2\kappa\theta > \sigma^2$, then it is a positive and stationary process.

We use the transition density to simulate the sample paths of the model (4.1) (See Cox, Ingersoll and Ross, 1985). The initial interest rate X_{t_0} at initial time t_0 is generated from the invariant density

of process (4.1), which is a Gamma distribution given by

$$p(z) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} z^{\alpha-1} e^{-y/\beta}$$

where $\alpha = 2\kappa\theta/\sigma^2$ and $\beta = \sigma^2/(2\kappa)$. Given the current interest rate $X_t = x$ at time t, $2cX_s$ at time $s > t$ is a noncentral chi-square conditional distribution with degrees of freedom $2q + 2$ and the noncentrality parameter $2u$ where

$$q = \tfrac{2\kappa\theta}{\sigma^2} - 1, \quad u = cxe^{-\kappa(s-t)}, \quad \text{and} \quad c = \tfrac{2\kappa}{\sigma^2(1-e^{-\kappa(s-t)})}.$$

We sampled the process at a weekly frequency with an interval $\Delta = 1/52$. The values of other parameters (κ, θ, σ) are cited from the work of Chapman and Pearson (2000) in our implementation, i.e. $\kappa = 0.21459$, $\theta = 0.08571$, $\sigma = 0.07830$. We generate a sample path of 1000 and replicate the experiments 100 times. We take $a = 0.05$, $b = 0.12$, and $s = t + \Delta$ for one-step forecasting in case (a) and $s = t + 2\Delta$ for two-step forecasting in case (b). The values of conditional density $g(y|x)$ are estimated at the observed sample points. Table 4 shows the simulation results and Figure 4 presents typical estimates of conditional densities. The RMSEs for estimating conditional density in two-step prediction are smaller than those for the one-step forecast. This is somewhat surprising but understandable. The conditional density for one-step forecasting tends to be larger (less spread) and hence has a larger estimation error in absolute terms. The estimates for the conditional density at $x = 0.085$ and $x = 0.12$ are reasonable, though there are not many local data points available for estimating the conditional density. The marginal density of X_t is Gamma(0.6, 0.0143) with mean 0.0857 and SD 0.0111. Thus, there are even fewer data points at the points around $x = 0.05$ which makes estimates unreliable.

Table 4: Summary of RMSEs for Example 4.

	case (a)				case (b)			
	CV	HWY	FYT	HY	CV	HWY	FYT	HY
mean	1.2088	1.2081	1.6531	2.1260	0.7259	0.7207	0.9828	1.4740
SD	0.0688	0.0652	0.1527	0.0589	0.0431	0.0431	0.1104	0.0297
median	1.2151	1.2146	1.6932	2.1198	0.7320	0.7267	0.9190	1.4715
robust SD	0.0393	0.0304	0.2109	0.0382	0.0354	0.0333	0.1504	0.0278

NOTE: Mean, SD, median and robust SD(=IQR/1.35) of RMSE for each method.

The performance of the HWY and CV methods is competitive. Note that the noncentral chi-square distribution with a large degree of freedom (which is 12 for the given parameters) is similar to a normal distribution (see Figure 4). Some researchers even use normal distribution to generate the CIR model, although this generating method incurs the problem of discretization errors. Hence, the CV method works as well as with the HWY method.

5 Applications

5.1 *Case study with Canadian lynx data*

This data set consists of 114 yearly observations on the proxy of the annual number of lynxs that were trapped in the Mackenzie River district of northwest Canada between the year 1821 and 1934. Details and thorough analysis of the data can be found in the book by Tong (1990). We take $a = 4$ and $b = 8.5$ and estimated the conditional density function of X_{t+1} given X_t for one-step forecasting using the CV approach as the bandwidth selection rule. The estimated conditional density function is given in Figure 5. It clearly contains many nonlinear features, such as bimodality.

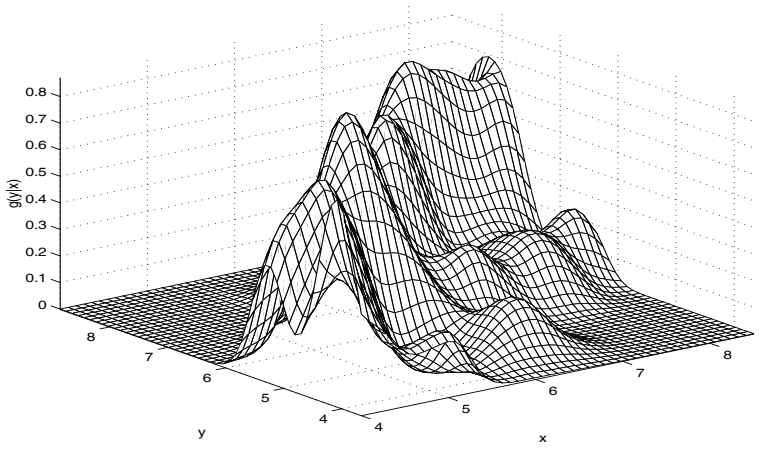

Figure 5: Estimated conditional density of X_t given $X_{t-1} = x$ for the Canadian lynx data using CV approach.

To check on the performance of the above three bandwidth selection rules, we use the data between 1821 and 1924 to estimate the conditional density function $\hat{g}(y|x)$ for the one-step forecasting, and use the last 10 data points to check the 90% predictive interval. All of the predictive intervals contain the corresponding true values. The average lengths of the predictive intervals using the CV, HWY and FYT approaches are 2.98, 3.31, and 3.69 respectively. The shortest average length is obtained by the CV approach. The method of estimating conditional density can also be applied to the two-step forecasting of X_{t+2} given X_t based on the CV, HWY, and FYT approaches. The average lengths of 90%-predictive intervals based on the CV, HWY and FYT approaches are 4.35, 4.56 and 4.53, respectively, which are broader than the one-step forecasting. Again, the CV approach provides, on average, the shortest predictive intervals.

5.2 *Case study with U.S. twelve-month treasury bill data*

This data set concerns the yields of U.S. twelve-month treasury bills from the secondary market rates. The data consist of 2112 weekly observations from July 17, 1959 to December 31, 1999.

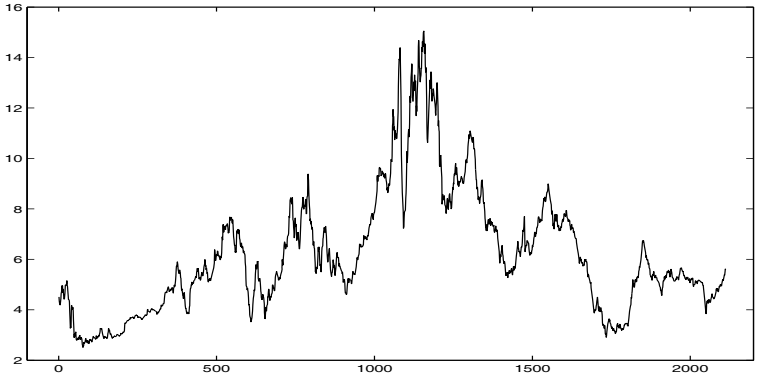

Figure 6: Time series for the yields of treasury bills from July 17, 1959 to December 31, 1999.

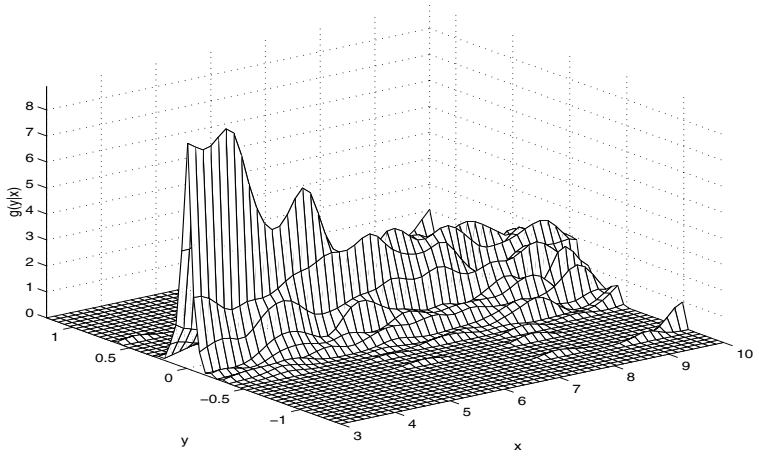

Figure 7: Estimated conditional density of X_t given $X_{t-1} = x$ for the 12-month Treasury bill data using the CV approach.

The time series plot is depicted in Figure 6. We take $a = 3$ and $b = 10$ and let $Y_t = Z_t - Z_{t-1}$, $X_t = Z_{t-1}$, where Z_t is the yields of twelve-month treasury bills. The estimated conditional density of Y_t given $X_t = x$ using the CV approach as the bandwidth selection rule is shown in Figure 7. A distinctive feature is that the conditional variance increases as the interest rates increase. The discrepancies between this empirically estimated transition density and that from the CIR model (Figure 4a) can be seen.

To ascertain on the performance, we use the first 1381 observations to estimate the conditional density of Y_t given $X_t = x$, and the last 14 years' observations to check the 95% predictive interval. A more interesting comparison is to construct a 95% conditional lower confidence limit based on the estimated conditional density. This lower confidence limit is related to the Value-at-Risk (VaR), a measure of risk of a portfolio in risk management (see Jorion, 2000).

Table 5 summarizes the average lengths and the coverage probabilities of the predictive intervals using the CV approach. Also included in Table 5 is the RiskMetrics of J.P. Morgan (1996), which is a popular method with which to forecast VaR. Let $r_t = Y_t/X_t$ be the observed return at time t. The idea behind the RiskMetrics is to estimate the volatility $\hat{\sigma}_t$ by exponential smoothing (see e.g.

Table 5: Performance comparisons of the RiskMetrics and double-kernel local linear regression using the CV approach.

		90% PI		95% lower bound	
Period		AL[a]	ECP[b]	ALB[c]	ER
1/1/1986–31/12/1999	RiskMetrics	0.33	91.57%	-0.16	5.61%
	CV	0.46	95.62%	-0.22	2.46%
1/1/1993–31/12/1999	RiskMetrics	0.27	91.80%	-0.14	3.83%
	CV	0.26	89.09%	-0.14	3.28%
1/8/1997–31/12/1999	RiskMetrics	0.24	89.76%	-0.12	7.09%
	CV	0.27	93.70%	-0.13	4.72%

[a]Averate length; [b]Empirical cover probability; [c]Average lower bound.

Gijbels, Pope, Wand, 1999)

$$\hat{\sigma}_t^2 = 0.94\hat{\sigma}_{t-1}^2 + 0.06r_{t-1}^2.$$

The 95% lower bound of r_t of RiskMetrics is $-1.645\sigma_t$. Hence, the 95% lower bound of Y_t by the RiskMetrics is $-1.645X_t\sigma_t$. The exceedence ratio (ER) is given by

$$\text{ER} = n^{-1}\sum_{t=T+1}^{T+n} I(r_t < \widehat{\text{VaR}}),$$

where $T+1$ and $T+n$ are the first and last observations in the validation period. It measures the performance of different VaR methods. Overall, our method tends to be more conservative, leading to high empirical coverage probability and low ER. Note that the RiskMetrics method is based on time-domain smoothing, which use mainly the recent data and the conditional density approach is based on state-domain smoothing, which uses mainly the historical data. In fact, our method does not use the data in the last 14 years. This can be improved by using a window of data close to the predicted time point, resulting in a time-varying prediction rule. An interesting challenge will be determining how to use both pieces of valuable information from both time-domain and state-domain smoothing to enhance the predictability.

To examine the impact of the period under consideration, we now use the data until December 31, 1992 as a training set and those from January 1, 1993 to December 31, 1999 as an out-sample period. In addition, we also employ the data until July 31, 1997 as a training period and the data after July 31 as a prediction period. The results for these two periods are also summarized in Table 5. They indicate clearly that the performance of each method depends on the training and prediction periods. The performance of the state-domain smoothing approach gets improved as the training period becomes longer so that more recent data are used in the prediction and more data are used in the estimation. On the other hand, because of its domain-smoothing, the estimation

of the RiskMetrics mainly uses the information in the past year, no matter how long the training period is.

Acknowledgments

Fan was partially supported by NSF grant DMS-0204329, a direct allocation RGC grant from the Chinese University of Hong Kong, and The Institute of Mathematical Sciences, CUHK.

Appendix

We now outline the key idea of the proof. Throughout the following proof, we use C to denote a generic constant, which may vary from line to line.

Proof of (3.5). We first compute the difference between $\hat{g}_h(y|x)$ and $\hat{g}_{h,-i}(y|x)$. To this end, we add the subscript "-i" to any quantities that do not involve the i-th data point (X_i, Y_i).

We first compare the difference between $s_{n,j}(x)$ and $s_{n,j,-i}(x)$. To facilitate the notation, we denote

$$W_{j,h_1}(z) = z^j W(z/h_1)/h_1.$$

Then,

$$s_{n,j,-i}(x) = (n-1)^{-1} \sum_{k \neq i} W_{j,h_1}(X_k - x).$$

By simple algebra,

$$s_{n,j,-i}(x) - s_{n,j}(x) = \frac{1}{n(n-1)} \sum_{k \neq i} W_{j,h_1}(X_k - x) - \frac{1}{n} W_{j,h_1}(X_i - x).$$

As W has a bounded support and is bounded, it follows that

$$|W_{j,h_1}(z)| \leq C h_1^{j-1},$$

and hence

$$|s_{n,j,-i}(x) - s_{n,j}(x)| \leq \frac{C h_1^{j-1}}{n}.$$

Substituting this into the definition of the equivalent kernel, and again using the fact that W has a bounded support, we can show easily that

$$|W_n(z;x) - W_{n,-i}(z;x)| \leq \frac{C}{n h_1},$$

for all z and x so that

$$s_{n,0,-i}(x) s_{n,2,-i}(x) - s_{n,1,-i}^2(x) > C^{-1} h_1^2.$$

The above holds with probability tending to one. Hence,

$$|W_n(z;x) - W_{n,-i}(z;x)| \leq O_P\left(\frac{C}{nh_1}\right). \tag{A.1}$$

Note that the above quantities involve only the design points. Hence, O_P-term will be exchangeable with the conditional expectation E_X, and for simplicity, we drop the notation O_P in (A.1). As $W(z)$ vanishes with, say, $|z| \geq 1$, it follows that

$$|W_n(z;x) - W_{n,-i}(z;x)| \leq \frac{C}{nh_1} I(|z| \leq 1). \tag{A.2}$$

We now investigate the difference between $\hat{g}_h(y|x)$ and $\hat{g}_{h,-i}(y|x)$. Observe that

$$|\hat{g}_{h,-i} - \hat{g}_h| \leq I_1 + I_2 + I_3, \tag{A.3}$$

where

$$I_1 = \frac{1}{nh_1h_2}\sum_{k\neq i}\left|W_{n,-i}\left(\frac{X_k - x}{h_1}; x\right) - W_n\left(\frac{X_k - x}{h_1}; x\right)\right| K\left(\frac{Y_k - y}{h_2}\right),$$

$$I_2 = \frac{1}{n(n-1)h_1h_2}\sum_{k\neq i}\left|W_{n,-i}\left(\frac{X_k - x}{h_1}; x\right)\right| K\left(\frac{Y_k - y}{h_2}\right),$$

and

$$I_3 = \frac{1}{nh_1h_2}\left|W_n\left(\frac{X_i - x}{h_1}; x\right)\right| K\left(\frac{Y_i - y}{h_2}\right).$$

We now deal with each of the above terms. By (A.2), we have

$$I_1 \leq \frac{1}{nh_1h_2}\sum_{k\neq i}\frac{C}{nh_1} I(|X_k - x| \leq h_1) K\left(\frac{Y_k - y}{h_2}\right).$$

By simple calculation,

$$E_X(I_1) \leq \frac{C}{n^2h_1^2}\sum_{k\neq i} I(|X_k - x| \leq h_1) = O_P\left(\frac{1}{nh_1}\right).$$

Similarly,

$$E_X(I_2) \leq \frac{C}{n(n-1)h_1}\sum_{k\neq i}\left|W_{n,-i}\left(\frac{X_k - x}{h_1}; x\right)\right|.$$

Note that by the Cauchy-Schwartz inequality, we have

$$\begin{aligned}
&\frac{1}{(n-1)h_1}\sum_{k\neq i}\left|W_{n,-i}\left(\frac{X_k - x}{h_1}; x\right)\right| \\
\leq\ & \frac{2s_{n,0,-i}(x)s_{n,2,-i}(x)}{s_{n,0,-i}(x)s_{n,2,-i}(x) - s^2_{n,1,-i}(x)} \\
=\ & 2 + o_P(1).
\end{aligned}$$

Consequently,

$$E_X(I_2) = O_P(\frac{1}{n}).$$

For I_3, we have

$$E_X(I_3) \leq \frac{C}{nh_1} \left| W_n\left(\frac{X_i - x}{h_1}; x\right)\right| = O_P\left(\frac{1}{nh_1}\right).$$

By (A.3), we have

$$E_X|\hat{g}_{h,-i}(y|x) - \hat{g}_h(y|x)| = O_P\left(\frac{1}{nh_1}\right). \tag{A.4}$$

We are now ready to prove (3.5). As $\hat{g}_{h,-i}(y|x)$ does not involve the i-th data point, by the double expectation formula, we have

$$E_X \sum_{i=1}^{n} \hat{g}_{h,-i}(Y_i|X_i) I(X_i \in [a,b]) = \sum_{i=1}^{n} \int E_X \hat{g}_{h,-i}(y|x) I(x \in [a,b]) g(y|x) f(x) dx dy.$$

Therefore, by (A.4), we have

$$E_X \frac{1}{n} \sum_{i=1}^{n} \hat{g}_{h,-i}(Y_i, X_i) I(X_i \in [a,b]) = \int E_X \hat{g}_h(y|x) I(x \in [a,b]) g(y|x) f(x) dx dy + O_P\left(\frac{1}{nh_1}\right).$$

This completes the proof of (3.5).

Proof of (3.6). Recall the definition $W_{j,h_1} = z^j W(z/h_1)/h_1$. We first note that, by Chebyshev's inequality,

$$W_{j,h_1}(X_i - x) = h_1^j \left\{ \int u^j W(u)[f(x) + h_1 u f'(x)] du + O_P\left(h_1^2 + \frac{1}{\sqrt{nh_1}}\right)\right\}.$$

Then, we compare the difference between $s_{n,j,-i}(x)s_{n,k,-i}(x)$ and $s_{n,j}(x)s_{n,k}(x)$. By simple algebra,

$$\begin{aligned}
&s_{n,j,-i}(x)s_{n,k,-i}(x) - s_{n,j}(x)s_{n,k}(x) \\
&= \sum_{p\neq i}\sum_{q\neq i} \frac{2n-1}{n^2(n-1)^2} W_{j,h_1}(X_p - x) W_{k,h_1}(X_q - x) \\
&\quad - \sum_{p=1}^{n} n^{-2} W_{j,h_1}(X_p - x) W_{k,h_1}(X_i - x) \\
&\quad - \sum_{q=1}^{n} n^{-2} W_{j,h_1}(X_i - x) W_{k,h_1}(X_q - x) + n^{-2} W_{j,h_1}(X_i - x) W_{k,h_1}(X_i - x).
\end{aligned}$$

As W has a bounded support and is bounded, it follows that

$$|s_{n,j,-i}(x)s_{n,k,-i}(x) - s_{n,j}(x)s_{n,k}(x)| \leq \begin{cases} \frac{Ch_1^{j+k}}{n}, & \text{both } j,k = 2. \\ \frac{Ch_1^{j+k+1}}{n}, & \text{either } j \text{ or } k = 1. \\ \frac{Ch_1^{j+k+2}}{n}, & \text{both } j,k = 1. \end{cases}$$

Substituting this into the definition of the equivalent kernel, and using the fact that W has a bounded support and is bounded, one can show that

$$|W_{n,-i}(z_1;x) W_{n,-i}(z_2;x) - W_n(z_1;x) W_n(z_2;x)| \leq \frac{C}{n} I(z_1 \leq h_1) I(z_2 \leq h_1),$$

for all z_1, z_2, and x such that

$$\{s_{n,0,-i}(x)s_{n_2,-i}(x) - s^2_{n,1,-i}(x)\}^2 \geq C^{-1}h_1^4.$$

We now investigate the difference between $\int \hat{g}_h(y|x)^2 dy$ and $\int \hat{g}_{h,-i}(y|x)^2 dy$. Observe that

$$\left| \int \hat{g}_{h,-i}(y|x)^2 dy - \int \hat{g}_h(y|x)^2 dy \right| \leq I_1 + I_2 + I_3 + I_4, \tag{A.5}$$

where

$$\begin{aligned} I_1 &= \sum_{k\neq i}\sum_{l\neq i} \frac{1}{(n-1)^2 h_1^2 h_2} \left| W_{n,-i}\left(\frac{X_k - x}{h_1}; x\right) W_{n,-i}\left(\frac{X_l - x}{h_1}; x\right) \right. \\ &\quad \left. - W_n\left(\frac{X_k - x}{h_1}; x\right) W_n\left(\frac{X_l - x}{h_1}; x\right) \right| K * K\left(\frac{Y_k - Y_l}{h_2}\right), \end{aligned}$$

$$\begin{aligned} I_2 = \sum_{k\neq i}\sum_{l\neq i} \frac{2n-1}{n^2(n-1)^2 h_1^2 h_2} &\left| W_{n,-i}\left(\frac{X_k - x}{h_1}; x\right) \right| \\ &\times \left| W_{n,-i}\left(\frac{X_l - x}{h_1}; x\right) \right| K * K\left(\frac{Y_k - Y_l}{h_2}\right), \end{aligned}$$

$$I_3 = 2\sum_{k=1}^{n} \frac{1}{n^2 h_1^2 h_2} \left| W_n\left(\frac{X_k - x}{h_1}; x\right) W_n\left(\frac{X_i - x}{h_1}; x\right) \right| K * K\left(\frac{Y_k - Y_i}{h_2}\right),$$

and

$$I_4 = \frac{1}{n^2 h_1^2 h_2} \left| W_n\left(\frac{X_i - x}{h_1}; x\right) \right|^2 K * K(0).$$

We now deal with each of the above terms. By simple calculation,

$$\begin{aligned} E_X(I_1) &\leq \frac{1}{(n-1)^2 h_1^2} \sum_{k\neq i}\sum_{l\neq i} \frac{C}{n} I(|X_k - x| \leq h_1) I(|X_l - x| \leq h_1) \\ &= O_P\left(\frac{1}{n}\right). \end{aligned}$$

Using the Cauchy-Schwartz inequality, we have

$$E_X(I_2) \leq O_P\left(\frac{1}{n}\right).$$

Similarly,

$$E_X(I_3) \leq O_P\left(\frac{1}{nh_1}\right),$$

and

$$E_X(I_4) \leq O_P\left(\frac{1}{n^2 h_1^2}\right).$$

By (A.5), we have

$$E_X \left| \int \hat{g}_{h,-i}(y|x)^2 dy - \int \hat{g}_h(y|x)^2 dy \right| = O_P\left(\frac{1}{nh_1}\right). \tag{A.6}$$

We are now ready to prove (3.6). Since $\int \hat{g}_{h,-i}(y|x)^2 dy$ does not involve the i-th data point, by the double expectation formula, we have

$$E_X \sum_{i=1}^{n} I(X_i \in [a,b]) \int \hat{g}_{h,-i}(y|X_i)^2 dy =$$
$$E_X \int \sum_{i=1}^{n} I(x \in [a,b]) \hat{g}_{h,-i}(y|x)^2 dy f(x) dx.$$

Therefore, by (A.6), we have

$$E_X \frac{1}{n} \sum_{i=1}^{n} I(X_i \in [a,b]) \int \hat{g}_{h,-i}(y|X_i)^2 dy =$$
$$E_X \int \hat{g}_h(y|x)^2 f(x) I(x \in [a,b]) dx dy + O_P\left(\frac{1}{nh_1}\right).$$

This completes the proof of (3.6).

References

Aït-Sahalia, Y. (1999), Transition densities for interest rate and other nonlinear diffusions. *The Journal of Finance*, **50**, 1361-1395.

Black, F., & Scholes, M. (1973), The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, **81**, 637-659.

Bashtannyk, D. M. and Hyndman, R.J. (2001). Bandwidth selection for kernel conditional density estimation. *Comp. Statist. Data Anal.*, **36**, 279-298.

Bowman, A. W. (1984), An Alternative Method of Cross-Validation for the Smoothing of Density Estimates. *Biometrika*, **71**, 353-360.

Chan, K. C., Karolyi, A. G., Longstaff, F. A. & Sanders, A. B. (1992), An Empirical Comparison of Alternative Models of the Short-term Interest Rate. *Journal of Finance*, **47**, 1209-1227.

Chan, K. S., & Tong, H. (1994), A Note on Noisy Chaos. *J. R. Statist. Soc.* B, **56** 301-311.

Chan, K.S. & Tong, H. (2001). *Chaos: A Statistical Perspective*. Springer, New York.

Chapman, D.A. and Pearson, N.D. (2000). Is the short rate drift actually nonlinear? *Journal of Finance*, **LV**, 355-388.

Cox, J. C., Ingersoll, J. E., & Ross, S. A. (1980), An Analysis of Variable Rate Loan Contracts. *Journal of Finance*, **35**, 389-403.

Cox, J. C., Ingersoll, J. E., & Ross, S. A. (1985), A Theory of the Term Structure of Interest Rates. *Econometrica*, **53**, 385-407.

Fan, J. (2003). A selective overview of nonparametric methods in financial econometrics. Submitted. *Research report 2003-03, Institute of Mathematical Sciences, Chinese University of Hong Kong.*

Fan, J., & Gijbels, I. (1995), Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation. *Journal of the Royal Statistical Society,* Ser. B, **57**, 371-394.

Fan, J., & Yao, Q. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods,* Springer-Verlag, New York.

Fan, J., Yao, Q., & Tong, H. (1996), Estimation of Conditional Densities and Sensitivity Measures in Nonlinear Dynamical Systems. *Biometrika*, **83**, 189-206.

Gijbels, I., Pope, A., and Wand, M.P. (1999), Understanding exponential smoothing via kernel regression, *Journal of the Royal Statistical Society, Series B*, **61**, 39–50.

Hall, P., Wolff, R. C. L., & Yao, Q. (1999), Methods for Estimating a Conditional Distribution Function. *Journal of the American Statistical Association*, **94**, 154-163.

Hansen, L.P., Scheinkman, J.A. and Touzi, N. (1998). Spectral methods for identifying scalar diffusions. *Journal of Econometrics*, **86**, 1-32.

Hyndman, R. J. and Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *Nonpara. Statist.*, **14**, 259-278.

Hyndman, R.J., Bashtannyk, D.M. and Grunwald, G.K. (1996). Estimating and visualizing conditional densities. *Jour. Comput. Graph. Statist.*, **5**, 315-336.

J.P. Morgan (1996), *RiskMetrics Technical Document.* Fourth edition, New York.

Jorion, P. (2000) *Value at Risk: The new benchmark for managing financial risk* (2nd ed.). McGraw-Hill, New York.

Polonik, W. and Yao, Q. (2000). Conditional minimum volume predictive regions for stochastic processes. *J. Ameri. Statist. Assoc*, **95**, 509-519.

Robinson, P. M. (1991), Consistent Nonparametric Entropy-Based Testing. *Rev. Econ. Studies*, **58**, 437-453.

Rosenblatt, M. (1969), Conditional Probability Density and Regression Estimators. In *Multivariate Analysis II,* Ed. P. R. Krishnaiah, pp. 25-31. New York: Academic Press.

Rudemo, M. (1982), Empirical Choice of Histograms and Kernel Density Estimators. *Scand. J. Statist.*, **9**, 65-78.

Ruppert, D. (1997), Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation. *Journal of the American Statistical Association*, **92**, 1049-1062.

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995), An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of the American Statistical Association*, **90**, 1257-1269.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Stone, M. (1974), Cross-Validatory Choice and Assessment of Statistical Predictions (with discussion). *Journal of the Royal Statistical Society*, Ser. B, **36**, 111-147.

Tjøstheim, D. (1994), Non-Linear Time Series: A Selective Review. *Scand. J. Statist*, **21**, 97-130.

Tong, H. (1990). *Non-Linear Time Series: A Dynamical System Approach*. Oxford University Press, Oxford.

Vasicek, O. A. (1977), An Equilibrium Characterization of the Term Structure. *Journal of Financial Economics*, **5**, 177-188.

Wand, M.P. & Jones, M.C. (1995), *Kernel Smoothing*, Chapman and Hall, London.