Biometrika (2004), **91,** 4, *pp.* 819–834
© 2004 Biometrika Trust
Printed in Great Britain

A crossvalidation method for estimating conditional densities

By JIANQING FAN and TSZ HO YIM

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

jfan@sta.cuhk.edu.hk   th_yim@sparc20c.sta.cuhk.edu.hk

Summary

We extend the idea of crossvalidation to choose the smoothing parameters of the 'double-kernel' local linear regression for estimating a conditional density. Our selection rule optimises the estimated conditional density function by minimising the integrated squared error. We also discuss three other bandwidth selection rules, an ad hoc method used by Fan et al. (1996), a bootstrap method of Hall et al. (1999) for bandwidth selection in the estimation of conditional distribution functions, modified by Bashtannyk & Hyndman (2001) to cover conditional density functions, and finally a simple approach proposed by Hyndman & Yao (2002). The performance of the new approach is compared with these three methods by simulation studies, and our method performs outstandingly well. The method is illustrated by an application to estimating the transition density and the Value-at-Risk of treasury-bill data.

Some key words: Bandwidth selection; Bootstrap; Conditional density function; Crossvalidation; Diffusion process; Financial application; Transition density.

1. Introduction

A conditional density provides the most informative summary of the relationship between independent and dependent variables. In particular, in stationary time series with Markovian structures, conditional densities characterise the probabilistic aspect of the time series. They determine, except for the initial distribution, the likelihood function of an observed time series and facilitate statistical prediction (Tong, 1990; Fan & Yao, 2003), as well as the study of nonlinear phenomena, such as the symmetry, multimodality structure and sensitivity measures of a time series (Chan & Tong, 2001).

The conditional density function plays a pivotal role in financial econometrics. It is directly related to the pricing formula of financial derivatives and inferences about parameters in financial models (Aït-Sahalia, 1999). Consider the continuous time model in which an economic variable, X_t, satisfies the 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 model is widely used in finance and economics, and includes many well-known single-factor models, such as those of Black & Scholes (1973), Vasicek (1977), Cox et al. (1980, 1985) and Chan et al. (1992), for modelling stock prices or interest rates. These models, except for the last, admit closed forms for the conditional probability density function, which allows us to evaluate explicitly the prices of financial derivatives based on X_t. For other cases, approximations of conditional densities are needed (Aït-Sahalia, 1999).

Nonparametric estimation of the drift and diffusion functions in (1·1) based on discretely observed data, with a fixed sampling frequency, is a challenging problem of high current interest in financial econometrics and statistics. An overview of and references for nonparametric techniques in financial economics are provided in an unpublished technical report from the Chinese University of Hong Kong by J. Fan. The transition density allows one to estimate the unknown functions from model (1·1). In fact, it determines the drift function $\mu(.)$ and diffusion function $\sigma(.)$ in model (1·1); see for example Hansen et al. (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 validity of the aforementioned well-known financial models.

A vast variety of papers use estimators of conditional densities as building blocks. Such papers include those of Robinson (1991), Tjøstheim (1994) and Polonik & Yao (2000), among others. However, in all of those papers, the conditional density function was estimated indirectly. Hyndman et al. (1996) studied the kernel estimator of a conditional density and its bias-corrected version. Fan et al. (1996) developed a direct estimation method via an innovative 'double-kernel' local linear approach. Despite its importance in various applications, the problem of estimating the conditional density function automatically has not been systematically studied since the pioneering work of Rosenblatt (1969). In particular, to our knowledge, no consistent data-driven procedure has been proposed for choosing smoothing parameters.

Fan et al. (1996) used a simple method for selecting the smoothing parameters. Hall et al. (1999) suggested, for a related topic, a bandwidth selection method for estimating a conditional distribution function together with a bootstrap approach. Bashtannyk & Hyndman (2001) and Hyndman & Yao (2002) proposed several simple and useful rules for selecting bandwidths for conditional density estimation. Hall et al. (2004) applied the crossvalidation technique to estimate the conditional density and to select relevant variables. The goal of the present paper is to develop a consistent data-driven bandwidth selection rule for estimating conditional density functions. The rule is based on the cross-validation method. Although this paper is motivated by the problem of time series data, we present our method in a more general context.

2. Estimation of conditional density

We assume that 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 et al. (1996).

Estimation of the conditional density can be regarded as a nonparametric regression problem. To make this connection, Fan et al. (1996) observed that, as $h_2 \to 0$,

$$E\{K_{h_2}(Y-y)|X=x\} \to g(y|x), \tag{2·1}$$

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 $X=x$. For each given x and y, the principle of local linear regression (Fan, 1992) suggests the minimisation 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 $\mathscr{X}$ be the design matrix of the local least-squares problem (2·2) and let

$$\mathscr{W}=\operatorname{diag}\{W_{h_1}(X_1-x),\ldots,W_{h_1}(X_n-x)\},$$
$$\mathscr{Y}=(K_{h_2}(Y_1-y),\ldots,K_{h_2}(Y_n-y))^{\mathrm{T}}.$$

We define

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

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

$$\hat{g}_h(y|x)=e_1^{\mathrm{T}}S_n^{-1}(x)T_n(x,y), \tag{2·3}$$

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

It is also instructive to express the estimated conditional density in the form of the equivalent kernel (Fan & Yao, 2003, p. 235). Let

$$W_n(z;x)=W(z)\frac{s_{n,2}(x)-zh_1 s_{n,1}(x)}{s_{n,0}(x)s_{n,2}(x)-s_{n,1}(x)^2}, \tag{2·4}$$

where $s_{n,j}(x)=n^{-1}\sum_{i=1}^{n}(X_i-x)^j W_{h_1}(X_i-x)$, for $j=0,1,2$. Then the estimator (2·3) can be written as

$$\hat{g}_h(y|x)=\frac{1}{nh_1h_2}\sum_{i=1}^{n}W_n\left(\frac{X_i-x}{h_1};x\right)K\left(\frac{Y_i-y}{h_2}\right). \tag{2·5}$$

3. Bandwidth selection

3·1. *Two simple methods*

Fan et al. (1996) proposed an ad hoc method for selecting the smoothing parameters. In the density estimation setting, the normal referencing rule (Silverman, 1986, p. 45) selects the bandwidth

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

where s_y is the sample standard deviation of Y.

For given bandwidth h_2 and y, (2·2) corresponds to nonparametric regression of $K_{h_2}(Y_i-y)$ on X_i. There are many data-driven methods for selecting the bandwidth h_1. These include crossvalidation (Stone, 1974), the residual-square criterion (Fan & Gijbels, 1995), the pre-asymptotic substitution method (Fan & Gijbels, 1995), the plug-in method

(Ruppert et al., 1995) and the empirical bias method (Ruppert, 1997), among others. We will refer to this method for selecting the smoothing parameters h_1 and h_2 as the Fan et al. approach. It is only a simple method, and is not expected to work in all situations. In our numerical implementations, we use the plug-in method of Ruppert et al. (1995) for choosing h_1. As shown below, the Fan et al. approach tends to oversmooth the conditional density in the y-direction.

Bashtannyk & Hyndman (2001) proposed an even simpler method based on the assumption that the conditional 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 Fan et al. (1996). Hyndman & Yao (2002) introduced new conditional density estimators based on local parametric modelling and derived several bandwidth selectors. When W and K are Gaussian kernels, Hyndman & Yao (2002) suggested using

$$h_1 = 0{\cdot}935(v\sigma^5/n|d_1|^5)^{1/6}, \quad h_2 = |d_1|h_1,$$

by assuming that the conditional density is $N(d_0 + d_1 x, \sigma^2)$ and the marginal density of X is $N(\mu, v^2)$. We will refer to this approach as the Hyndman–Yao method. We include this method for comparison because of its attractiveness in implementation, which makes it extremely useful for getting an initial idea of the bandwidths to be used. In fact, Hyndman & Yao (2002) use this method only as an initial estimator for their more effective method.

3·2. *Bootstrap bandwidth selection*

Hall et al. (1999) suggested a bootstrap approach for selecting the smoothing parameters in the context of estimating conditional distribution functions. Their idea can be adapted here to the conditional density function context; see Bashtannyk & Hyndman (2001). First, we fit a simple parametric model

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

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

$$Y_i^* = \hat{a}_0 + \hat{a}_1 X_i + \ldots + \hat{a}_{\hat{k}} X_i^{\hat{k}} + \hat{\sigma}\varepsilon_i^*. \tag{3·2}$$

Hence, we obtain a bootstrap sample of $\{Y_1^*, \ldots, Y_n^*\}$ and a bootstrap estimate, $\hat{g}_h^*(y|x)$. Let

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

be the bootstrap estimator of the absolute deviation error of $\breve{g}(y|x)$; the expectation is taken with respect to the bootstrap sample. Finally, we choose h to minimise

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

where $[a, b]$ is an interval on 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 Hall et al. approach.

3·3. *A crossvalidation method*

The above two approaches are simple and ad hoc. They are not intended to optimise the estimated conditional densities. The bootstrap method provides a 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, and then the bootstrap method fails to select the optimal bandwidths. We here extend a crossvalidation idea of Rudemo (1982) and Bowman (1984).

Let $f(x)$ be the marginal density of $\{X_i\}$ and let $[a, b]$ be an interval on which we wish to estimate the conditional density. Define the integrated squared error as

$$\begin{aligned}\textsc{ise} &= \int \{\hat{g}_h(y|x) - g(y|x)\}^2 f(x) I(x \in [a, b]) dx\, dy \\ &= \int \hat{g}_h(y|x)^2 f(x) I(x \in [a, b]) dx\, dy - 2 \int \hat{g}_h(y|x) g(y|x) f(x) I(x \in [a, b]) dx\, dy \\ &\quad + \int g(y|x)^2 f(x) I(x \in [a, b]) dx\, dy. \qquad (3{\cdot}3)\end{aligned}$$

Note that the last term does not depend on h and can be ignored in minimisation of ISE with respect to h.

A reasonable estimator of (3·3) is

$$\textsc{cv}(h) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \in [a, b]) \int \hat{g}_{h,-i}(y|X_i)^2 dy - \frac{2}{n} \sum_{i=1}^{n} \hat{g}_{h,-i}(Y_i|X_i) I(X_i \in [a, b]), \quad (3{\cdot}4)$$

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 calculated explicitly by using (2·5). If we define ISE as

$$n^{-1} \sum_{i=1}^{n} I(X_i \in [a, b]) \int \{\hat{g}_h(y|X_i) - g(y|X_i)\}^2 dy,$$

then the first term of the quadratic expansion is known and does not need to be estimated. Its corresponding crossvalidation function is given by

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

The crossvalidation method can also be used to select the bandwidth of the Hyndman & Yao (2002) estimator for conditional density. Further applications of the crossvalidation method, including selection of relevant variables in conditional densities, can be found in Hall et al. (2004).

Estimating conditional density is much more involved than the density estimation setting and 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 or not the proposed crossvalidation method is reasonable. To appreciate how much 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 because of the random denominator. Instead, we can compute the conditional expectation. However, for the times series applications, the design points $\{X_t, t = 1, \ldots, T-1\}$ involve nearly the 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 & 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 consider the independent random sample setting.

Let $\{(X_i, Y_i), i=1, \ldots, n\}$ be an independent 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 $X_1, \ldots, X_n$, namely $E_X(Z)=E(Z|X_1, \ldots, X_n)$. 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 $f(.)$ is continuous and positive in an open interval containing $[a, b]$ and $g(.|x)$ is bounded for x in an open interval containing $[a, b]$,*

$$E_X \frac{1}{n} \sum_{i=1}^{n} \hat{g}_{h,-i}(Y_i|X_i) I(X_i \in [a, b]) = E_X \int \hat{g}_h(y|x) g(y|x) f(x) I(x \in [a, b]) dx\, dy + O_P\left(\frac{1}{nh_1}\right). \tag{3·5}$$

Furthermore, under the additional condition that $f'(.)$ 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 $O_P\{1/(nh_1h_2)\}$ (Fan et al., 1996).

4. SIMULATION STUDIES

We consider simulation studies to evaluate and compare bandwidth selection methods for estimating the conditional density that were described in § 3. For each simulation, the performance of the selection rule is evaluated by the root mean squared error,

$$\text{RMSE} = \frac{[\sum_i \{\hat{g}_h(y_i|x_i) - g(y_i|x_i)\}^2 I(x_i \in [a, b])]^{\frac{1}{2}}}{\sum_i I(x_i \in [a, b])},$$

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 and scale models*). We consider a simple quadratic model

$$Y_i = 0{·}23 X_i(16 - X_i) + 0{·}4\varepsilon_i \quad (i \geqslant 1). \tag{4·1}$$

The noise term is simulated from the following situations:

(i) $\{\varepsilon_i\}$ are independent standard normal random variables,

(ii) $\{\varepsilon_i\}$ are a random sample from the t_2 distribution, and

(iii) $\{\varepsilon_i\}$ are independent and follow the t_4 distribution.

For cases (i)–(iii), X_i are independent Un$[0, 16]$. We also consider the following time series setting:

(iv) $X_{i+1} = Y_i$ with some initial value X_1, where the noise variables ε_i are independent random variables having the same distribution as the sum of 48 independent random variables each uniformly distributed on $[-0{·}25, 0{·}25]$.

In case (iv), the bounded support of ε_i is necessary for the stationarity of the time series (Chan & Tong, 1994). The conditional density of this model was studied by Fan et al. (1996).

For each of the 100 samples of size $n = 1000$, we calculate the root mean squared errors 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 (i)–(iii), and $a = 4$ and $b = 14$ for case (iv).

We summarise the results in Table 1. The means, standard derivations, medians and robust standard deviations of RMSE for each of the three bandwidth selection rules are given; the robust standard deviation is equal to the interquartile range divided by 1·35. In general, the Hyndman–Yao (2002) and Fan et al. (1996) approaches produce larger values of RMSE than those of crossvalidation and the Hall et al. (1999) method. The performance of both crossvalidation and the Hall et al. method are comparable. Note that we only employ the simple rule of Hyndman & Yao (2002) for ease of implementation. It is expected that their more sophisticated method should be more effective than what we present here. Cases (i) and (iv) are ideal for the Hall et al. approach because this is the model in which the bootstrap sample was generated. Nevertheless, the proposed crossvalidation approach works comparably to the Hall et al. method, which is a parametric approach for this model. However, the performance of the Hall et al. approach deteriorates when the tails of the noise ε_i become heavier; see case (ii) and Fig. 1, which presents a typical estimated conditional density with median performance. In fact, for case (ii), the crossvalidation method outperforms the other two approaches substantially.

The improvement of the crossvalidation method can be even more substantial when the regression function is not quadratic, as in

$$Y_i = 20 \cos\left(\frac{\pi X_i}{10}\right) + \varepsilon_i, \qquad (4{\cdot}2)$$

where the noise ε_i are independent standard normal random variables and

(I) X_i are independent Un$[-20, 20]$ random variables;

(II) $X_{i+1} = Y_i$ with some initial value X_1, as studied by Fan et al. (1996).

Table 1: *Example* 1. *Summary of the root mean squared error for model* (4·1) *showing*, *for each method*, *the mean* (*standard deviation*) *in the first row*, *and the median* (*robust standard deviation*) *of the root mean squared error* ($\times 10^{-3}$) *in the second row*

	Case (i)	Case (ii)	Case (iii)	Case (iv)
CV	1·0899 (0·0601)	0·7641 (0·1497)	1·0143 (0·0683)	1·0903 (0·0673)
	1·0882 (0·0616)	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)

CV, crossvalidation method; HWY, Hall et al. (1999) method; FYT, Fan et al. (1996) method; HY, Hyndman–Yao (2002) method.

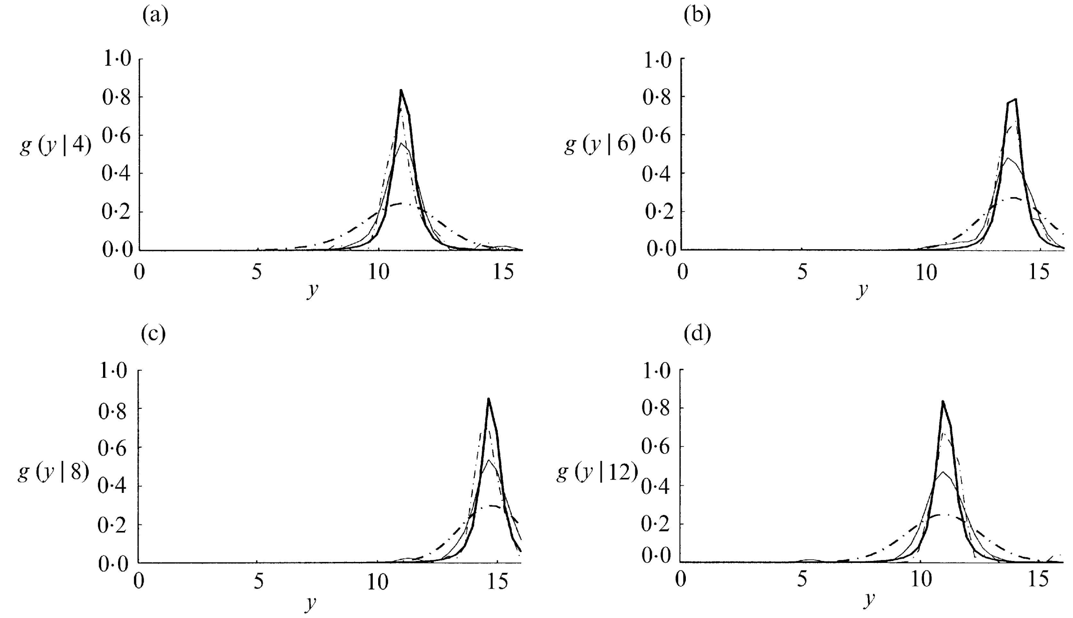

Fig. 1: Example 1. Estimated conditional densities for model (4·1) with case (ii). Estimated conditional densities for (a) $x = 4$, (b) $x = 6$, (c) $x = 8$ and (d) $x = 12$, using the crossvalidation method (thin dash-dot curve), Hyndman–Yao (2002) method (thin solid curve), and Fan et al. (1996) method (thick dash-dot curve) approaches, are compared with the true densities (thick solid curve).

Table 2 summarises the results based on 100 simulations and sample size $n = 1000$ with $a = -18$ and $b = 18$ for case (i), and $a = -17$ and $b = 20$ for case (ii). Model (4·2) deviates from the bootstrap model and this is the main reason for the poor performance of the Hall et al. method. This shows that the Hall et al. method is not consistent.

We have also tested the four bandwidth selection methods on the following scale model,

$$Y_i|X_i \sim \mathrm{Ga}(3, X_i^2 + 1), \qquad (4·3)$$

where X_i are independent standard uniform random variables. The crossvalidation and the Hyndman–Yao methods perform comparably, and outperform the Hall et al. and Fan et al. methods by approximately 10% in terms of RMSE. This shows that the simple method of Hyndman & Yao can also perform very well beyond its assumed class of models.

Table 2: *Example* 1. *Summary of the root mean squared error for model* (4·2), *showing the mean, standard deviation, median and robust standard deviation* ($\times 10^{-3}$) *for each method*

	Case (i)				Case (ii)			
	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

CV, crossvalidation method; HWY, Hall et al. (1999) method; FYT, Fan et al. (1996) method; HY, Hyndman–Yao (2002) method; SD, standard deviation.

Example 2 (Cox–Ingersoll–Ross model). We consider the well-known Cox–Ingersoll–Ross model (Cox et al., 1985) for the evolution of interest rates:

$$dX_t = \kappa(\theta - X_t)dt + \sigma(X_t)^{\frac{1}{2}} dW_t \quad (t \geqslant t_0). \tag{4·4}$$

This 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 model (4·4); see Cox et al. (1985). The interest rate X_{t_0} at initial time t_0 is generated from the invariant density of process (4·4), which is a Gamma distribution given by $p(z) = \{\Gamma(\alpha)\beta^{\alpha}\}^{-1} z^{\alpha-1} e^{-y/\beta}$, 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 $s > t$ has a noncentral chi-squared conditional distribution with $2q + 2$ degrees of freedom and noncentrality parameter $2u$, where

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

We sampled the process at a weekly frequency with an interval $\Delta = \frac{1}{52}$. The values of other parameters (κ, θ, σ) are cited from the work of Chapman & Pearson (2000) in our implementation, that is $\kappa = 0·21459$, $\theta = 0·08571$ and $\sigma = 0·07830$. 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 (i) and $s = t + 2\Delta$ for two-step forecasting in case (ii). The values of the conditional density $g(y|x)$ are estimated at the observed sample points. Table 3 shows the simulation results and Fig. 2 presents typical estimates of conditional densities. The root mean squared errors 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, i.e. 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. The marginal density of X_t is Ga(0·6, 0·0143), with mean 0·0857 and standard deviation 0·0111. Thus, there are even fewer data points around $x = 0·05$, which makes estimates unreliable.

The performance of the Hall et al. (1999) and crossvalidation methods is competitive. Note that the noncentral chi-squared distribution with many degrees of freedom, 12 for the given parameters, is similar to a normal distribution; see Fig. 2. Some researchers

Table 3. *Summary of root mean squared errors for Example* 2, *showing the mean, standard deviation, median and robust standard deviation of* RMSE *for each method*

	Case (i)				Case (ii)			
	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

CV, crossvalidation method; HWY, Hall et al. (1999) method; FYT, Fan et al. (1996) method; HY, Hyndman–Yao (2002) method; SD, standard deviation.

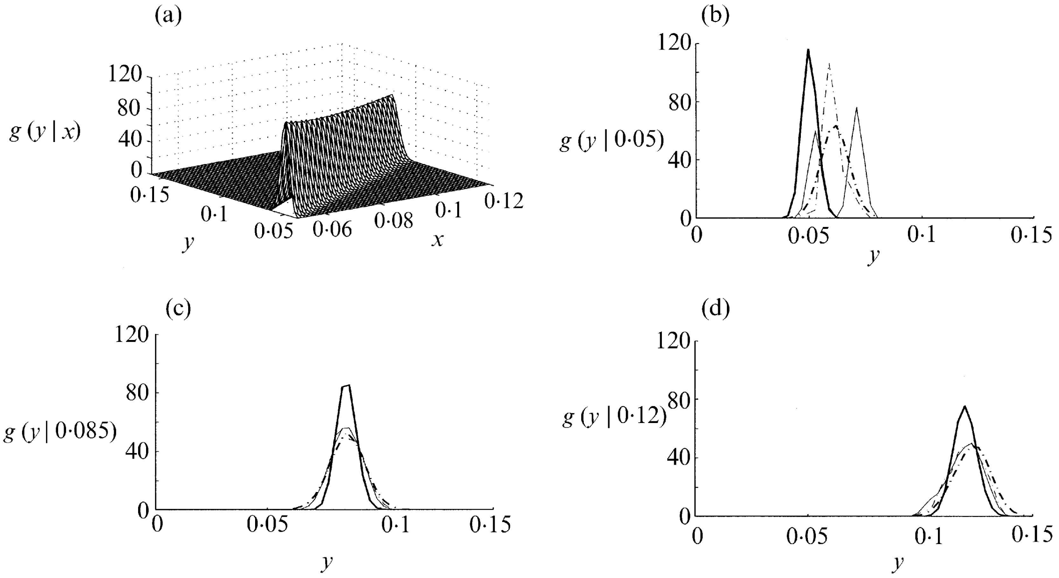

Fig. 2: Example 2. Two-step transition densities for the Cox–Ingersoll–Ross (1985) model. (a) shows the true conditional density function. In (b)–(d) estimated densities, for (b) $x=0{\cdot}05$, (c) $x=0{\cdot}085$ and (d) $x=0{\cdot}12$, using the crossvalidation method (thin dash-dot curve), Hall et al. (1999) method (thin solid curve) and Fan et al. (1996) method (thick dash-dot curve) approaches, are compared with the true densities (thick solid curve).

even use a normal distribution to generate the Cox–Ingersoll–Ross model, although this generating method incurs the problem of discretisation errors. Hence, the crossvalidation method works as well as the Hall et al. method.

5. APPLICATIONS

Our illustration concerns the yields of the U.S. twelve-month treasury bills from the secondary market rates. The data consist of 2112 weekly observations from 17 July 1959 to 31 December 1999. The time series plot is depicted in Fig. 3(a). We take $a=3$ and $b=10$ and let $Y_t=Z_t-Z_{t-1}$ and $X_t=Z_{t-1}$, where Z_t is the yield of twelve-month treasury bills. The estimated conditional density of Y_t given $X_t=x$, using the crossvalidation approach as the bandwidth selection rule, is shown in Fig. 3. 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 Cox–Ingersoll–Ross (1985) model, see Fig. 2(a), can be seen.

To investigate 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 90% 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 (Jorion, 2000).

Table 4 summarises the average lengths and the coverage probabilities of the predictive intervals using the crossvalidation approach. Also included in Table 4 is the RiskMetrics approach of J. P. Morgan, given in the firm's 'RiskMetrics technical document', and which

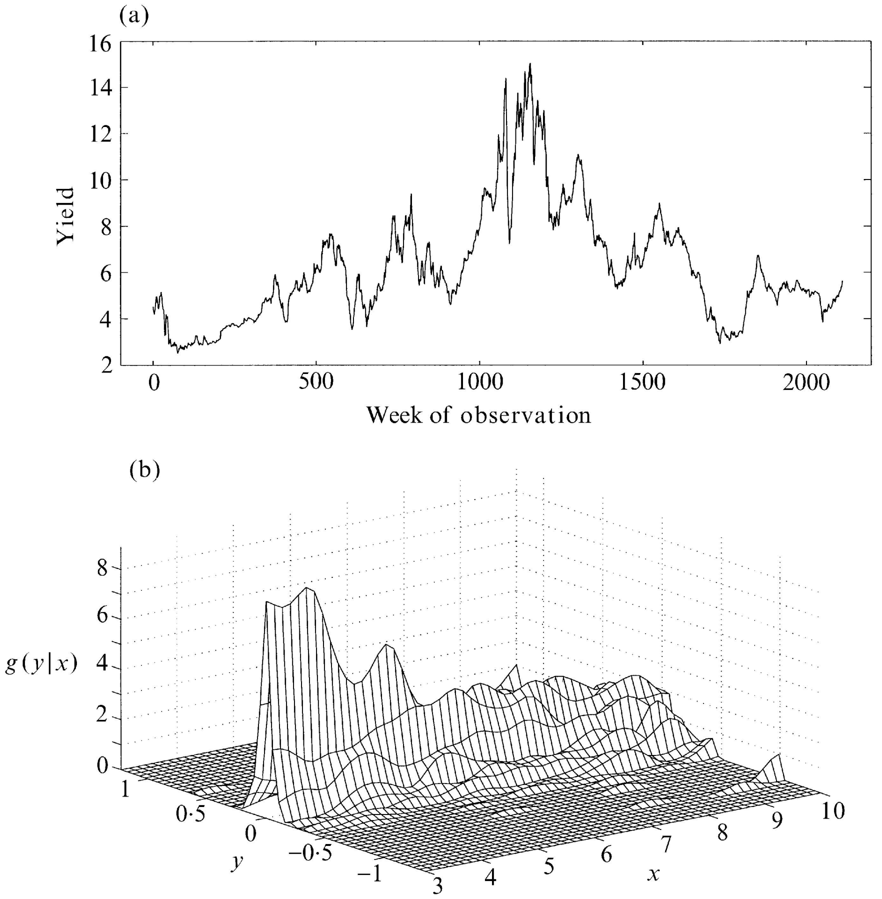

Fig. 3. Yields (%) of treasury bills from 17 July 1959 to 31 December 1999. (a) shows the data; (b) shows the estimated conditional density of X_t given $X_{t-1}=x$, based on the crossvalidation approach.

Table 4: *Treasury bill data. Performance comparisons of the RiskMetrics and double-kernel local linear regression using the crossvalidation approaches*

		90% PI		95% lower bound	
Period		AL	ECP	ALB	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%

Methods: RiskMetrics of J. P. Morgan; CV, crossvalidation.
PI, predictive interval; AL, average length; ECP, empirical coverage probability; ALB, average lower bound; ER exceedence ratio.

is a popular method with which to forecast Value-at-Risk. 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:

$$\hat{\sigma}_t^2 = 0{\cdot}94\hat{\sigma}_{t-1}^2 + 0{\cdot}06 r_{t-1}^2.$$

The 95% lower bound of r_t of RiskMetrics is $-1{\cdot}645\sigma_t$; that is, the J. P. Morgan estimate of the 95%VaR at time t is $-1{\cdot}645\sigma_t$. The exceedence ratio is given by

$$\text{ER} = n^{-1} \sum_{t=T+1}^{T+n} I(r_t < -1{\cdot}645\sigma_t),$$

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 uses mainly recent data, and the conditional density approach is based on state-domain smoothing, which uses mainly historical data. In fact, our method does not use data from the most recent 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 to determine how to use information from both time-domain smoothing and state-domain smoothing to enhance the predictability.

To examine the impact of the period under consideration, we now use the data up until 31 December 1992 as a training set and those from 1 January 1993 to 31 December 1999 as a prediction period. In addition, we also employ the data up until 31 July 1997 as a training period and the data after 31 July as a prediction period. The results for these two periods are also summarised in Table 4. They indicate clearly that the performance of each method depends on the training and prediction periods. The performance of the state-domain smoothing approach improves 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 time-domain smoothing, the RiskMetrics estimation mainly uses the information from the most recent year, no matter how long the training period is.

Acknowledgement

Fan was partially supported by the U.S. National Science Foundation, by a direct allocation of the Research Grants Council grant from the Chinese University of Hong Kong and by The Institute of Mathematical Sciences of the Chinese University of Hong Kong. The authors acknowledge gratefully the constructive comments by the editor, an associate editor and a referee, and assistance of Dr Heng Peng.

Appendix

Proofs

We now outline the key ideas of the proofs. Throughout, 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 ith data point (X_i, Y_i).

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

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

By simple algebra, we obtain

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

As W has a bounded support and is bounded, it follows that $|W_{j,h_1}(z)| \leqslant C h_1^{j-1}$, and hence that

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

Substituting this into the definition of the equivalent kernel, we can show easily that

$$|W_n(z; x) - W_{n,-i}(z; x)| \leqslant \frac{C}{nh_1},$$

for all z and x such 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,

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

Note that the above quantities involve only the design points. Hence, the 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, $|z| \geqslant 1$, it follows that

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

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

$$|\hat{g}_{h,-i} - \hat{g}_h| \leqslant I_1 + I_2 + I_3, \qquad (A{\cdot}3)$$

where

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

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

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

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

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

By simple calculation, we have

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

Similarly,

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

Note that, by the Cauchy–Schwartz inequality, we have

$$\frac{1}{(n-1)h_1}\sum_{k\neq i}\left|W_{n,-i}\left(\frac{X_k-x}{h_1};x\right)\right|\leqslant 2+o_P(1).$$

Consequently, $E_X(I_2)=O_P(n^{-1})$. For I_3, we have

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

It follows from (A·3) that

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

We are now ready to prove (3·5). As $\hat{g}_{h,-i}(y|x)$ does not involve the ith data point, by the double expectation formula, we have

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

Therefore, it follows from (A·4) that

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

This completes the proof of (3·5).

Proof of (3·6). We first note that, by Chebyshev's inequality,

$$s_{n,j}(x)=h_1^j\left[\int u^jW(u)\{f(x)+h_1uf'(x)\}du+O_P\left\{h_1^2+\frac{1}{\surd(nh_1)}\right\}\right].$$

Next we calculate the difference between $s_{n,j,-i}(x)s_{n,k,-i}(x)$ and $s_{n,j}(x)s_{n,k}(x)$. Using the fact that W has a bounded support and is bounded, we obtain

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

Substituting this into the definition of the equivalent kernel, one can show that

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

for all z_1, z_2 and x such that

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

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

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

where

$$I_1 = \sum_{k \neq i} \sum_{l \neq i} \frac{1}{(n-1)^2 h_1^2 h_2} \left| W_{n,-i}\left(\frac{X_k - x}{h_1}; x\right) W_{n,-i}\left(\frac{X_l - x}{h_1}; x\right) \right.$$

$$\left. - W_n\left(\frac{X_k - x}{h_1}; x\right) W_n\left(\frac{X_l - x}{h_1}; x\right) \right| K*K\left(\frac{Y_k - Y_l}{h_2}\right),$$

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

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

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

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

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

By the Cauchy–Schwartz inequality, we can show that

$$E_X(I_2) \leqslant O_P\left(\frac{1}{n}\right), \quad E_X(I_3) \leqslant O_P\left(\frac{1}{nh_1}\right), \quad E_X(I_4) \leqslant O_P\left(\frac{1}{n^2 h_1^2}\right).$$

By (A·5), we have

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

We are now ready to prove (3·6). Since $\int \hat{g}_{h,-i}(y|x)^2 dy$ does not involve the ith 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 obtain (3·6).

References

Aït-Sahalia, Y. (1999). Transition densities for interest rate and other non-linear diffusions. *J. Finance* **54**, 1361–95.

Bashtannyk, D. M. & Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Comp. Statist. Data Anal.* **36**, 279–98.

Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities. *J. Polit. Econ.* **81**, 637–59.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353–60.

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. *J. Finance* **47**, 1209–27.

Chan, K. S. & Tong, H. (1994). A note on noisy chaos. *J. R. Statist. Soc.* B **56**, 301–11.

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

Chapman, D. A. & Pearson, N. D. (2000). Is the short rate drift actually nonlinear? *J. Finance* **55**, 355–88.

Cox, J. C., Ingersoll, J. E. & Ross, S. A. (1980). An analysis of variable rate loan contracts. *J. 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. (1992). Design-adaptive nonparametric regression. *J. Am. Statist. Assoc.* **87**, 998–1004.

Fan, J. & Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. R. Statist. Soc.* B **57**, 371–94.

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

Fan, J., Yao, Q. & Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83**, 189–206.

Hall, P., Racine, J. & Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *J. Am. Statist. Assoc.* To appear.

Hall, P., Wolff, R. C. L. & Yao, Q. (1999). Methods for estimating a conditional distribution function. *J. Am. Statist. Assoc.* **94**, 154–63.

Hansen, L. P., Scheinkman, J. A. & Touzi, N. (1998). Spectral methods for identifying scalar diffusions. *J. Economet.* **86**, 1–32.

Hyndman, R. J. & Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparam. Statist.* **14**, 259–78.

Hyndman, R. J., Bashtannyk, D. M. & Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *J. Comp. Graph. Statist.* **5**, 315–36.

Jorion, P. (2000). *Value at Risk: The New Benchmark for Managing Financial Risk*, 2nd ed. New York: McGraw-Hill.

Polonik, W. & Yao, Q. (2000). Conditional minimum volume predictive regions for stochastic processes. *J. Am. Statist. Assoc.* **95**, 509–19.

Robinson, P. M. (1991). Consistent nonparametric entropy-based testing. *Rev. Econ. Studies* **58**, 437–53.

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. *J. Am. Statist. Assoc.* **92**, 1049–62.

Ruppert, D., Sheather, S. J. & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Am. Statist. Assoc.* **90**, 1257–69.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with Discussion). *J. R. Statist. Soc.* B **36**, 111–47.

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.

Vasicek, O. A. (1977). An equilibrium characterization of the term structure. *J. Finan. Econ.* **5**, 177–88.

[*Received June* 2003. *Revised April* 2004]