Conditional Minimum Volume Predictive Regions For Stochastic Processes *

Wolfgang Polonik
Division of Statistics
University of California at Davis
Davis, CA 95616, USA

Qiwei Yao
Institute of Mathematics and Statistics
University of Kent at Canterbury
Canterbury, Kent CT2 7NF, UK

Abstract

Motivated by interval/region prediction in nonlinear time series, we propose a minimum volume predictor (MV-predictor) for a strictly stationary process. The MV-predictor varies with respect to the current position in the state space and has the minimum Lebesgue measure among all regions with the nominal coverage probability. We have established consistency, convergence rates, and asymptotic normality for both coverage probability and Lebesgue measure of the estimated MV-predictor under the assumption that the observations are from a strong mixing process. Applications with both real and simulated data sets illustrate the proposed methods.

Keywords: Conditional distribution, level set, minimum volume predictor, Nadaraya-Watson estimator, nonlinear time series, predictor, strong mixing.

*Supported partially by EPSRC Grants and an EU HCM Program. The authors thank Professor Peter C Young for making available the rainfall-flow data analyzed in §4. The helpful comments of the editor, an associate editor and a reviewer are gratefully acknowledged.

1 Introduction

In any serious attempt to forecast, a point prediction is only a beginning. A predictive interval or, more generally, a predictive region is much more informative. In the context of linear time series models with normally distributed errors, the predictive distributions are normal. Therefore, the predictive intervals are easily obtained using mean plus and minus a multiple of the standard deviation. The width of such an interval, even for multi-step ahead prediction, is constant over the whole state space unless the model has conditional heteroscedastic noise. Predictive intervals constructed in this way have also been used in some special nonlinear models (*e.g.* threshold autoregressive models, see Tong and Moeanaddin 1988, Davies, Pemberton and Petruccelli 1988). However, the above method is no longer pertinent when the predictive distribution is not normal, which, unfortunately, is the case for most nonlinear time series models (Chan and Tong 1994). Recent studies on pointwise prediction of nonlinear time series have revealed that the prediction accuracy does depend on the current position in the state space (Yao and Tong 1994a, and the references therein). Yao and Tong (1995a) proposed to construct predictive intervals using conditional quantiles (percentiles) for nonlinear time series. However, interval predictors so constructed are not always appropriate when the predictive distributions are asymmetric or multi-modal. Asymmetric distributions have been widely used in modeling economic activities (Brännäs and De Gooijer 1992, and the references therein). Further, skewed predictive distributions may occur in multi-step ahead prediction even though the errors in the models have symmetric distributions. Multi-modal phenomena often indicate model uncertainty. The uncertainty may be caused by factors beyond the variables specified in the prediction. (See §2 below for some examples.) In order to cope with the possible skewness and multi-modality of the underlying predictive distribution, we propose to use a (conditional) minimum volume set, which we call the *(conditional) minimum volume predictor* (MV-predictor), among all the candidate regions in a given class (*e.g.* intervals). The MV-predictor depends on the current position in the state space. It forms a region where the predictive distribution has highest mass concentration, in the sense that it has minimal Lebesgue measure among all the sets in a given class with the nominal coverage probability. Especially, the MV-predictor of all the predictive intervals is the one with the shortest length. In fact, the MV-predictor has the aforementioned properties among *all* the sets with the nominal coverage probability, provided the model implied by the given class is correct. (See the discussion below Definition 2.2 in §2.2.) For a symmetric and unimodal predictive distribution, an MV-predictor

reduces to a quantile interval.

The minimum volume approach has a long history in the statistical literature, and a well-known example in its early time is the so-called *shorth* (Andrews *et al.* 1972). Additional literature on minimum volume sets can be found in Polonik (1997). A closely related concept *excess mass* was introduced independently by Hartigan (1987), and Müller and Sawitzki (1987,1991). See also Nolan (1991), and Polonik (1995). Hyndman (1995) seemed to be the first paper to use the minimum volume approach (under a different name) for time series prediction. Under the assumption that the predictive distribution is of a known parametric form, Hyndman (1995) estimated the MV-predictor based on a simulation method. However until now there has been a lack of appreciation of the general methodology and its properties, although empirical development of using it for time series prediction can be found in Yao and Tong (1995b), Hyndman (1996) and De Gooijer and Gannoun (1997).

In this paper, we establish various asymptotic properties of the predictive regions relying on the asymptotic results on conditional empirical processes reported in Polonik and Yao (1998). Under a general setting which admits time series modeling as a special case, we construct an estimator for the MV-predictor directly based on a Nadaraya-Watson estimator of the predictive distribution. Under the assumption that the observed data are from a strictly stationary and strong mixing process, we have established consistency for the estimated MV-predictor using an L_1-distance, and the asymptotic normality for both coverage probability and Lebesgue measure. We have also derived an explicit rate at which the estimated MV-predictor converges. Comparing with the results on the global (*i.e.* unconditional) minimum volume sets reported in Polonik (1997), the convergence rates for *conditional* MV-predictors are typically slower, although in similar form (see *e.g.* Remark 3.3(b) in §3 below). This reflects the distinction between local and global fittings. We also propose a bootstrap scheme to choose the state-dependent bandwidths for the purpose of estimating MV-predictors.

We use the Nadaraya-Watson estimator for the predictive distribution simply to keep the theory as simple as possible. It is conceivable that similar results hold if the predictive distribution is estimated by local linear regression (Tsybakov 1986, Fan 1992, Fan, Hu and Truong 1994, Yu and Jones 1998) or adjusted Nadaraya-Watson method (Hall, Wolff and Yao 1999).

The paper is organized as follows. §2 introduces the minimum volume predictor and its estimator. The advantages of using the proposed method are demonstrated through two simple time series models. We present all the theoretical results in §3. §4 reports the simulation results with a

nonlinear AR(2) model. The application to a rainfall-flow data from a catchment in Wales is also reported. The Appendix contains the key technical proofs.

2 Methodology

Suppose that X is an observable d-dimensional random vector, and Y is a d'-dimensional random vector which is unobservable. It is of interest to predict Y from X. In univariate time series context, $d' = 1$ and X typically denotes a vector of lagged values of Y. A predictive region $\Omega(\alpha|X) \subset \mathbf{R}^{d'}$ for Y from X with a nominal coverage probability $\alpha \in [0, 1]$ satisfies the condition

$$P\{\ Y \in \Omega(\alpha|x) \mid X = x\ \} \ \geq\ \alpha, \quad x \in \mathbf{R}^d. \tag{2.1}$$

Often we tend to choose $\Omega(\alpha|x)$ to be a connected set (*e.g.* an interval in the case that $d' = 1$). However, the consideration in accuracy of prediction leads us to search for a set which has the minimum volume (*i.e.* Lebesgue measure) among all the sets fulfilling condition (2.1). The resulting set is not necessarily connected. To illustrate the basic ideas of our approach, we give some numerical illustration via two toy models before the presentation of the formal definition of the minimum volume predictor.

2.1 Two toy models

We start with a simple quadratic model

$$Y_t = 0.23Y_{t-1}(16 - Y_{t-1}) + 0.4\epsilon_t, \tag{2.2}$$

where $\{\epsilon_t\}$ is a sequence of independent random variables each with the standard normal distribution truncated in the interval $[-12,\ 12]$. The conditional distribution of Y_t given $X_t = Y_{t-m}$ is symmetric for $m = 1$, but not necessarily so for $m > 1$. For example, the conditional density function at $X_t = 8$ with $m = 3$ is depicted in Fig.1(a), which is obviously skewed to the left. The curve was estimated using the 10,000 independent samples, each of them generated by iterating equation (2.2) three times with the starting point at 8. The kernel density estimator was used with Gaussian kernel and bandwidth 0.389. Based on this density function, two types of predictive intervals with three different coverage probabilities are specified in Table 1. The quantile interval $I(\alpha|x)$ is the interval with the $(0.5 - \alpha/2)$-th and the $(0.5 + \alpha/2)$-th percentiles as its two end-points. The minimum volume interval $M_1(\alpha|x)$ is the shortest interval among all the intervals

with coverage probability not smaller than α. For example, $I(\alpha|x) = [5.11, 14.96]$ for $\alpha = 0.95$ and $x = 8$. It contains some lower density points near its left end-point due to the skewness of the distribution; see Fig.1(a). The predictor $M_1(\alpha|x) = [6.50, 15.30]$ could be regarded as a *compressed* shift to the right of the quantile interval with 10.57% reduction in its length. Obviously, the accuracy of prediction has been substantially improved by using the minimum volume interval.

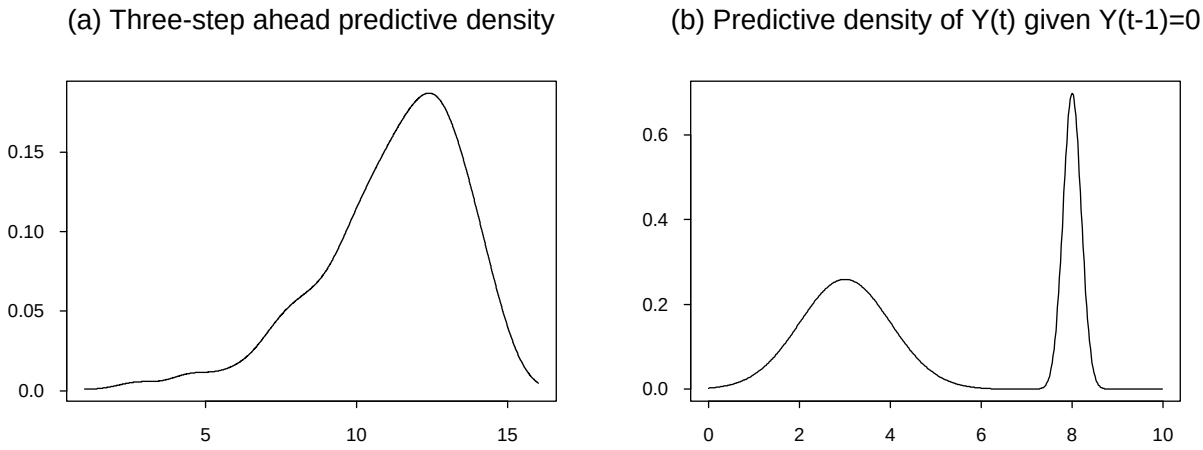

Figure 1: *(a) The conditional density function of Y_t given $Y_{t-3} = 8$ for model (2.2). (b) The conditional density function of Y_t given $Y_{t-1} = 0$ for model (2.3).*

Now we consider the model

$$Y_t = 3\cos\left(\frac{\pi Y_{t-1}}{10}\right) + Z_t + \frac{1}{0.8Z_t + 1}\epsilon_t, \tag{2.3}$$

where $\{\epsilon_t\}$ and $\{Z_t\}$ are two independent *i.i.d.* sequences with $\epsilon_1 \sim N(0,1)$ and $P(Z_1 = 0) = 0.65 = 1 - P(Z_1 = 5)$. For the sake of illustration, we assume that the 'exogenous' variable Z_t is unobservable. We predict Y_t from $X_t \equiv Y_{t-1}$ only. Thus the (theoretical) least square conditional point predictor is $3\cos(0.1\pi Y_{t-1}) + 1.75$, which is obviously not satisfactory. It is easy to see that the conditional distribution of Y_t given Y_{t-1} is a mixture of two normal distributions. Fig.1(b) depicts the conditional density function at $Y_{t-1} = 0$. For the three different values of α, Table 1 records the three types of predictive regions: the quantile interval $I(\alpha|x)$, the minimum volume intervals $M_1(\alpha|x)$, and the minimum volume region with at most two intervals $M_2(\alpha|x)$. We can see that the percentile interval fails to do a reasonable job simply because the predictive intervals are too wide. The improvement by using $M_1(\alpha|x)$ is not substantial unless the coverage probability α is small enough such that the probability mass within one mode exceeds α. The

region $M_2(\alpha|x)$ is much shorter in length (i.e. Lebesgue measure), and therefore offers a much more accurate prediction. All three $M_2(\alpha|x)'$s consist of two disconnected intervals, which clearly reveals the uncertainty in Y_t caused by the 'hidden' variable Z_t. The coverage probabilities of the two intervals centered at 3 and 8 are 0.61 and 0.34, 0.38 and 0.32, and 0.20 and 0.30 in order when the global coverage probability is 0.95, 0.70 and 0.50 respectively.

Table 1: Three predictive regions with coverage probability α: the quantile intervals $I(\alpha|x)$, the minimum volume interval $\mathrm{M}_1(\alpha|x)$, and the minimum volume region with at most two intervals $\mathrm{M}_2(\alpha|x)$. The percentage decreases in volume (*i.e.* Lebesgue measure) relative to $I(\alpha|x)$ are recorded in parentheses.

α	Predictor	Predictive regions for Y_t from Y_{t-3} at $Y_{t-3}=8$ for model (2.2)	Predictive regions for Y_t from Y_{t-1} at $Y_{t-1}=0$ for model (2.3)
0.95	$I(\alpha\|x)$	[5.11, 14.96]	[1.23, 8.30]
	$\mathrm{M}_1(\alpha\|x)$	[6.50, 15.30] (10.57%)	[1.50, 8.42] (2.12%)
	$\mathrm{M}_2(\alpha\|x)$		[1.15, 4.84]∪[7.54, 8.46] (34.79%)
0.70	$I(\alpha\|x)$	[8.66, 13.95]	[2.26, 8.04]
	$\mathrm{M}_1(\alpha\|x)$	[9.64, 14.15] (8.72%)	[2.80, 8.29] (5.02%)
	$\mathrm{M}_2(\alpha\|x)$		[2.18, 3.82]∪[7.66, 8.34] (59.86%)
0.50	$I(\alpha\|x)$	[9.83, 12.95]	[2.71, 7.89]
	$\mathrm{M}_1(\alpha\|x)$	[10.69, 13.56] (8.01%)	[1.80, 4.20] (53.67%)
	$\mathrm{M}_2(\alpha\|x)$		[2.60, 3.40]∪[7.71, 8.29] (73.36%)

In summary, we should seek for the minimum volume predictive regions when the conditional distribution of Y given X is skewed or/and multi-modal. The number of intervals used in the predictor should be equal, or at least close, to the number of modes of the conditional distribution, subject to practical feasibility.

2.2 Minimum volume predictive region

We use $F(\cdot|x)$ to denote the conditional distribution (*i.e.* the conditional measure) of Y given $X=x$, which is also called predictive distribution. We use the same F to denote the conditional distribution function in the sense that $F(y|x)=F((-\infty,y]|x)$. We use $g(\cdot|x)$ to denote its density function if it exists. We define the minimum volume predictor as follows.

Definition 2.1 *Let $\mathcal{C}$ denote a class of measurable subsets of $\mathbf{R}^{d'}$. Then any*

$$\mathrm{M}_{\mathcal{C}}(\alpha|x) \in \operatorname*{argmin}_{C\in\mathcal{C}}\{\mathrm{Leb}(C) : F(C|x)\geq\alpha\}, \quad \alpha\in[0,1], \ \ x\in\mathbf{R}^d, \tag{2.4}$$

is called a (conditional) minimum volume predictor (MV-predictor) in $\mathcal{C}$ at level α, *where* Leb(C) *denotes the Lebesgue measure of* C. *We denote its volume as*

$$\mu_{\mathcal{C}}(\alpha|x) = \text{Leb}(\text{M}_{\mathcal{C}}(\alpha|x)). \tag{2.5}$$

The above definition is based a prescribed class $\mathcal{C}$ which should contain all the interesting candidate predictors. The prior knowledge on the 'shape' of $F(\cdot|x)$ plays an important role in determining $\mathcal{C}$; see §2.1. On the other hand, $\mathcal{C}$ cannot contain, for example, all the subsets of $\mathbf{R}^{d'}$, which will make both theoretical exploration and practical implementation unnecessarily difficult. The MV-predictor is not always unique, and it may not even exist for some $\mathcal{C}$. However it exists and is unique if the density $g(.|x)$ exists and $g(\cdot|x) \in \mathcal{M}_{\mathcal{C}}(\alpha)$ which is defined as follows.

Definition 2.2 *For any probability density function* ϕ *on* $\mathbf{R}^{d'}$, *the* level set *at level* $\lambda > 0$ *is defined as* $\Gamma_\phi(\lambda) = \{y : \ \phi(y) \geq \lambda\}$. *For completeness we define* $\Gamma_\phi(0)$ *to be the support of* ϕ. *For* $\alpha \in [0,1]$ *define that* $\phi \in \mathcal{M}_{\mathcal{C}}(\alpha)$ *if and only if there exists a* $\lambda_\alpha > 0$ *for which*

$$\Gamma_\phi(\lambda_\alpha) \in \mathcal{C} \quad \textit{and} \quad \Phi(\Gamma_\phi(\lambda_\alpha)) = \alpha,$$

I have changed F **to** Φ **in the above formula – QY***???* *where* Φ *denotes the probability measure corresponding to* ϕ.

Under the assumption that $g(\cdot|x) \in \mathcal{M}_{\mathcal{C}}(\alpha)$, the set $\Gamma_{g(\cdot|x)}(\lambda_\alpha)$ is a MV-predictor (at level α), which is unique up to Leb-nullsets. In fact, this condition ensures that $\Gamma_{g(\cdot|x)}(\lambda_\alpha)$ has the smallest Lebesgue measure among *all* predictive regions with coverage probability at least α. Ideally the class $\mathcal{C}$ should be rich enough to contain $\Gamma_{g(\cdot|x)}(\lambda_\alpha)$ for all interesting values of α. In fact, selecting $\mathcal{C}$ can be viewed as selecting a statistical model, namely, the class of all the probability density functions with level sets in $\mathcal{C}$. (See also Remark 2.3 (b).) For a more thorough discussion about the modeling aspect, we refer to Polonik (1995, 1997).

Remark 2.3 We list some interesting features of MV-predictors in the case $d' = 1$ below. Let $\mathcal{I}_k$ denote the class of sets which are unions of at most k intervals. We write $\text{M}_{\mathcal{I}_k}(\alpha|x) = \text{M}_k(\alpha|x)$, $\mu_{\mathcal{I}_k}(\alpha|x) = \mu_k(\alpha|x)$ and $\mathcal{M}_{\mathcal{I}_k}(\alpha) = \mathcal{M}_k(\alpha)$.

(a) Suppose $g(\cdot|x)$ is symmetric and unimodal in the sense that there exists a point m_0 such that $g(\cdot|x)$ is strictly increasing to the left of m_0 and strictly decreasing to the right. Then $\text{M}_k(\alpha|x)$, for all $k \geq 1$, reduces to the quantile interval $I(\alpha|x) = [q(0.5 - \alpha/2|x),\ q(0.5 + \alpha/2|x)]$, where $q(\alpha|x)$ is the conditional quantile satisfying the equation $F(q(\alpha|x)|x) = \alpha$.

(b) If the conditional density $g(\cdot|x)$ is p-modal, then $g(\cdot|x) \in \mathcal{M}_p(\alpha)$. We should use the MV-predictor of $\mathcal{I}_k$ with $k = p$. On the other hand, we always tend to choose a small k in practice. In fact, the shortest predictive *interval* $\mathrm{M}_1(\alpha|x)$ is often appealing since it is a connected set.

(c) Similar to the unconditional *i.i.d.* case (Einmahl and Mason 1992, and Polonik 1997), the volume of the MV-predictor, $\mu_{\mathcal{C}}(\alpha|x)$, can be regarded as a generalized (conditional) quantile. To this end, we let $\mathcal{C}_\infty = \{(-\infty, y],\ y \in \mathbf{R}\}$ in (2.4), and replace the Lebesgue measure by the function ν defined as $\nu((-\infty, y]) = y$. Then the resulting 'MV-predictor' is $(-\infty, q(\alpha|x)]$ with the 'volume' $\nu((-\infty, q(\alpha|x)]) = q(\alpha|x)$.

2.3 Estimation

We assume that $\{(X_t, Y_t)\}$ is a strictly stationary process, and it has the same marginal distribution as $(X,\ Y) \in \mathbf{R}^d \times \mathbf{R}^{d'}$. Of interest is to estimate $\mathrm{M}_{\mathcal{C}}(\alpha|x)$ for a given class $\mathcal{C}$ based on observations $\{(X_t, Y_t), 1 \le t \le n\}$.

Note that for any given measurable set $C \subset \mathbf{R}^{d'}$, $E\{I_{(Y \in C)}|X = x\} = F(C|x)$. This regression relationship suggests the following Nadaraya-Watson estimator for the conditional distribution $F(\cdot|x)$:

$$\widehat{F}_n(C|x) = \sum_{t=1}^{n} I_{\{Y_t \in C\}} K\left(\frac{X_t - x}{h}\right) \Big/ \sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right),$$

where $K(\cdot) \ge 0$ is a kernel function on $\mathbf{R}^d$, and $h > 0$ is a bandwidth. $\widehat{F}_n(\cdot|x)$ is called the *empirical conditional distribution.*

Replacing $F(\cdot|x)$ by $\widehat{F}_n(\cdot|x)$ in (2.4), we obtain an estimator for the MV-predictor

$$\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x) \in \operatorname*{argmin}_{C \in \mathcal{C}} \{\mathrm{Leb}(C) : \widehat{F}_n(C|x) \ge \alpha\}. \tag{2.6}$$

We denote its volume and actual (unknown) coverage probability as

$$\widehat{\mu}_{\mathcal{C}}(\alpha|x) = \mathrm{Leb}(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x)), \quad \tilde{\alpha}_{\mathcal{C}} = F\{\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x)|x\}. \tag{2.7}$$

Remark 2.4 (a) Since we apply the kernel smoothing on a d-dimensional variable X in estimating $F(\cdot|x)$, the estimation suffers from the so-called 'curse-of-dimensionality', as all the other local estimation methods. For large d, it is an interesting and challenging problem to compress the information on Y contained in X into a lower dimensional variable before applying the proposed method, which is obviously beyond the scope of this paper.

(b) An obvious alternative to our minimum volume approach is to adopt a level set approach to construct a predictive region based on an estimated conditional density function (Fan, Yao and

Tong 1996). No need to specify $\mathcal{C}$ in the level set approach could be convenient; see, for example, Hyndman (1995). But we argue that there are added advantages to use the MV-predictor from a properly selected class $\mathcal{C}$. For example, we can always obtain a predictor consisting of at most k intervals with k prescribed by letting $\mathcal{C} = \mathcal{I}_k$. Further, the volumes of the MV-predictors with different k's provide valuable information on the shape of the predictive distribution; see Fig.4 in §4 below. Finally, the estimation for conditional density function involves local smoothing for $d+d'$ (instead of d) variables, which will further increase the difficulties associated with the 'curse of dimensionality'.

(c) Simple classes such as balls, rectangles are always appealing candidates for $\mathcal{C}$. In fact, more complicated is $\mathcal{C}$, the slower is the convergence rate of the estimator; see Theorem 3.2 and Remark 3.3(a) in §3 below.

2.4 Bootstrap bandwidth selector

Like all other kernel smoothers, the quality of our estimator depends crucially on the choice of the bandwidth h. However, the conventional data-driven bandwidth selectors such as cross-validation do not appear to have obvious analogues in the context of estimating MV-predictors. Deriving asymptotically optimal bandwidths is a tedious matter. Using plug-in methods requires explicit estimation of complex functions. Such complexity is arguably not justified, not least because the conditional measure $F(C|x)$ is often approximately monotone in x (*e.g.* x behaves like a location parameter) and so has only limited opportunity for complex behavior.

Instead, we suggest a bootstrap scheme to select the bandwidth, which is similar to the bandwidth selector for estimation of conditional distribution functions suggested by Hall, Wolff and Yao (1999). To simplify the presentation, we outline the scheme for the case that $d' = 1$ and $k = 1$ only. We fit a parametric model

$$Y_i = G(X_i, q) + \epsilon_i, \tag{2.8}$$

where $G(x, q)$ denotes a polynomial function of x and q is a set of indices indicating the terms included in G. We assume that $\{\epsilon_i\}$ are independent with a common distribution $N(0, \sigma^2)$ and $\sigma^2 > 0$ unknown. The parameters in G and σ^2 are estimated from the data. We form a parametric estimator $\check{\mathrm{M}}_1(\alpha|x)$ based on the above model. By Monte Carlo simulation from the model, we compute a bootstrap version $\{Y_1^*, \ldots, Y_n^*\}$ from (2.8) based on given observations $\{X_1, \ldots, X_n\}$, and with that a bootstrap version $\widehat{\mathrm{M}}_1^*(\alpha|x)$ of $\widehat{\mathrm{M}}_1(\alpha|x)$ with $\{(X_i, Y_i)\}$ replaced by $\{(X_i, Y_i^*)\}$.

Define

$$D(h) = E[\mathrm{Leb}\{\widehat{\mathrm{M}}_1^*(\alpha|x) \ \triangle \ \check{\mathrm{M}}_1(\alpha|x)\}|\{X_i, Y_i\}],$$

where $A \triangle B = (A - B) \cup (B - A)$ is the symmetric difference of sets A and B. Choose $h = \hat{h}$ to minimize $D(h)$.

In principle there are no difficulties to extend the above idea for estimation of $M_k(\alpha|x)$ with $k \geq 2$. We may, for example, choose the distribution of ϵ_t to be a mixture of k normal distributions. However, the bootstrap searching for $\widehat{\mathrm{M}}_k^*(\alpha|x)$ with $k \geq 2$ is computationally expensive if still feasible. In this paper, we simply use the bandwidth selected for $k = 1$ in the case of $k = 2$ as well. Our experience suggests that the choice between the two predictors $\widehat{\mathrm{M}}_1(\alpha|x)$ and $\widehat{\mathrm{M}}_2(\alpha|x)$ does not depend on the bandwidth sensitively unless a bifurcation occurs to the conditional distribution $F(\cdot|x)$ around x. Note that our problem is different in nature from that concerned in Silverman's test for multimodality (Silverman 1981). If we were interested in determining the number of modes in the curve $P(x) \equiv F(C|x)$, the bandwidth used in estimation would play a critical role.

3 Theoretical properties

In this section, we always assume that $x \in \mathbf{R}^d$ is given and $f(x) > 0$, where $f(\cdot)$ denotes the marginal density of X, and c denotes some generic constant, which may be different at different places. Furthermore, all stochastic quantities are assumed to be measurable, and we write $d_{F(\cdot|x)}(A, B) = F(A \Delta B|x)$.

Theorem 3.1 (Uniform consistency)
Suppose that $\mathrm{M}_{\mathcal{C}}(\alpha|x)$ *is uniquely defined (up to Leb-nullsets) by* (2.4)*, and the following two conditions hold.*

(i) $\mu_{\mathcal{C}}(\alpha|x)$ *is continuous in* $\alpha \in [a, b]$ *for some* $0 < a < b < 1$,

(ii) $\sup_{C \in \mathcal{C}} |\widehat{F}_n(C|x) - F(C|x)| \to 0$ *in probability (or almost surely) as* $n \to \infty$.

Then as $n \to \infty$,

$$\sup_{\alpha \in [a,b]} d_{F(\cdot|x)}(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x), \mathrm{M}_{\mathcal{C}}(\alpha|x)) \to 0 \quad \text{in probability (or almost surely).}$$

The above theorem is formulated in general terms. Condition (ii) is the key, which can be verified for $d' = 1$ and $\mathcal{C} = \mathcal{I}_k$ by appealing to the standard results of conditional empirical processes; see,

for example, Bosq (1996). Polonik and Yao (1998) justified the condition for $d' > 1$ and more general $\mathcal{C}$???.

All the results in this section could be formulated in a similar way, relying on certain properties of the set-indexed empirical conditional distribution. (See also Polonik 1997 for the unconditional *i.i.d.* case.) However, it is of interest to justify those properties under appropriate mixing conditions. Therefore, we present the results below in a more explicit manner. To this end, we introduce some regularity conditions first.

(A1) The kernel function K is bounded and symmetric, and $\lim_{u\to\infty} ||u||^d K(u) = 0$.

(A2) $f \in C_{2,d}(b)$, where $C_{2,d}(b)$ denotes the class of bounded real-valued functions with bounded second order partial derivatives.

(A3) $F(\cdot|x)$ has a Lebesgue-density $g(\cdot|x) \in C_{2,d'}(b)$. Moreover, for each $C \in \mathcal{C}$ we have $F(C|\cdot) \in C_{2,d}(b)$ such that $\sup_{C\in\mathcal{C}} \left|\frac{\partial^2}{\partial x_i \partial x_j} F(C|x)\right| < \infty \ \ \forall\, 1 \le i,j \le d$.

(A4) $||\int vv^T K(v)\, dv|| < \infty$.

(A5) The process $\{(X_t, Y_t)\}$ is strong mixing, *i.e.*

$$\alpha(j) = \sup_{A\in\mathcal{F}_{-\infty}^0,\, B\in\mathcal{F}_j^\infty} |P(AB) - P(A)P(B)| \to 0, \quad \text{as } j \to \infty, \tag{3.1}$$

where $\mathcal{F}_s^t$ denotes the σ-algebra generated by $\{(X_i, Y_i), s \le i \le t\}$. Further we assume that $\alpha(k) \le b\,\zeta^k$ for some $b > 0$ and $0 < \zeta < 1$.

(A6) For all $s \le t < q \le r$ the joint density function of (X_s, X_t, X_q, X_r) exists and is bounded by a constant independent of (s, t, q, r).

Now we introduce the notion of *metric entropy with bracketing* which provides a measure of the richness (or complexity) for a class $\mathcal{C}$. This notion is closely related to *covering numbers*. We adopt L_1-type covering numbers using the bracketing idea. The bracketing reduces to *inclusion* when it is applied to classes of sets rather than classes of functions. For each $\epsilon > 0$, the covering number is defined as

$$\begin{aligned} N_I(\epsilon, \mathcal{C}, F(\cdot|x)) = \inf\{n \ge 1 : \exists\, C_1, \ldots, C_n \in \mathcal{C} \text{ such that} \\ \forall\, C \in \mathcal{C} \ \ \exists\, 1 \le i,j \le n \text{ with } C_i \subset C \subset C_j \text{ and } F(C_j - C_i|x) < \epsilon\}. \end{aligned} \tag{3.2}$$

The quantity $\log N_I(\epsilon, \mathcal{C}, F(\cdot|x))$ is called the *metric entropy with inclusion* of $\mathcal{C}$ with respect to $F(\cdot|x)$. A pair of sets C_i, C_j with $C_i \subset C \subset C_j$ is called a *bracket* for C. Estimates for such

covering numbers are known for many classes. (See, *e.g.* Dudley 1974, 1984.) We will often assume below that either $\log N_I(\epsilon, \mathcal{C}, F(\cdot|x))$ or $N_I(\epsilon, \mathcal{C}, F(\cdot|x))$ behave like a power of ϵ^{-1}. We say that condition (R_γ) holds if

$$\log N_I(\epsilon, \mathcal{C}, F(\cdot|x)) < H_\gamma(\epsilon), \quad \text{for all } \epsilon > 0, \tag{R_γ}$$

where

$$H_\gamma(\epsilon) = \begin{cases} \log(A\,\epsilon^{-r}) & \text{if } \gamma = 0, \\ A\,\epsilon^{-\gamma} & \text{if } \gamma > 0, \end{cases} \tag{3.3}$$

for some constants $A, r > 0$. The large is γ, the richer (the more complicated) is the class $\mathcal{C}$.??? In fact (R_0) holds for intervals, rectangles, balls, ellipsoids, and for classes which are constructed from those by performing set operations union, intersection and complement finitely many times. Especially, the classes $\mathcal{I}_k$ used above fulfill (R_0). The classes of convex sets in $\mathbf{R}^d$ $(d \geq 2)$ fulfill condition (R_γ) with $\gamma = (d-1)/2$. Other classes of sets satisfying (R_γ) with $\gamma > 0$ can be found in Dudley (1974, 1984).

The following theorem concerns the rates of convergence of $d_{F(\cdot|x)}(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x), \mathrm{M}_{\mathcal{C}}(\alpha|x))$. This is the only place where we assume that the model implied by the class $\mathcal{C}$ is correct, *i.e.* $g(\cdot|x) \in \mathcal{M}_{\mathcal{C}}(\alpha)$. Under this condition, $d_{F(\cdot|x)}(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x), \mathrm{M}_{\mathcal{C}}(\alpha|x)) = d_{F(\cdot|x)}(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x), \Gamma_{g(\cdot|x)}(\lambda_\alpha))$.

Theorem 3.2 (Rates of convergence)
Let conditions (A1) – (A6), *and* (R_γ) *hold. Assume that for some* $\alpha \in (0,1)$, $g(\cdot|x) \in \mathcal{M}_{\mathcal{C}}(\alpha)$ *with corresponding level* $\lambda_\alpha > 0$. *Suppose further that* $\Gamma_{g(\cdot|x)}(\lambda)$ *is Lipschitz continuous in* λ *at* $\lambda = \lambda_\alpha$ *with respect to* $d_{F(\cdot|x)}$. *Suppose that*

$$|\widehat{F}_n(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x)|x) - F(\mathrm{M}_{\mathcal{C}}(\alpha|x)|x)| = o_P\left((nh^d)^{-1/2}\right). \tag{3.4}$$

Then for any $\eta > 0$ *and*

$$h = c\,\max\left(n^{-\frac{1}{d+(3+\gamma)}}, n^{-\frac{1}{d+2(3\gamma+1)}}\right),$$

we have that as $n \to \infty$,

$$d_{F(\cdot|x)}(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x), \Gamma_{g(\cdot|x)}(\lambda_\alpha)) = \begin{cases} o_P(n^{-\frac{1}{d+(3+\gamma)}+\eta}) & \text{if } \gamma < 1/5, \\ o_P(n^{-\frac{1}{d+2(3\gamma+1)}}) & \text{if } \gamma \geq 1/5. \end{cases}$$

Remark 3.3 (a) The convergence rate depends on the richness of the class $\mathcal{C}$, and is monotonically decreasing as γ (*i.e.* the richness) increasing.???

(b) For $\gamma = 0$, which includes the case $\mathcal{C} = \mathcal{I}_k$, the rate given in the above theorem is in fact of the form $(nh^d)^{-1/3+\eta}$. Note that the effective sample size in estimating *conditional* minimum volume set is nh^d instead of n. The above rate is in fact in alignment with the convergence rate $n^{-1/3+\eta}$ in estimating a *unconditional* minimum volume set with *i.i.d.* data (Polonik 1997). For the classes $\mathcal{I}_k$ it seems plausible that one can adapt the methods from Kim and Pollard (1990) to show that $(nh^d)^{-1/3}$ actually is the exact rate of convergence.

(c) In contrast to Theorems 3.4 and 3.6 below we cannot use the optimal bandwidth $h = O(n^{-\frac{1}{d+4}})$ here. The bandwidth h has to be even smaller to ensure the bias term $\sup_{C\in\mathcal{C}}(E\widehat{F}_n(C|x) - F(C|x))$, which is of order $\sqrt{nh^d}\ h^2$, tend to zero fast enough.

(d) It can be proved that assumption (3.4) is fulfilled for many classes $\mathcal{C}$, including the sets of balls, rectangles, ellipsoids, and convex sets in $\mathbf{R}^{d'}$. **Any possible references? –QY**??? We only give an heuristic explanation as follows. Note that the empirical conditional distribution $\widehat{F}_n(\cdot|x)$ has mass at the single point Y_t

$$\widehat{F}_n(\{Y_t\}|x) = K\left(\frac{X_t - x}{h}\right) / \sum_{t'=1}^{n} K\left(\frac{X_t' - x}{h}\right) \sim \frac{1}{nh^d} f^{-1}(x) K\left(\frac{X_t - x}{h}\right),$$

which is in the order of $1/(nh^d)$ since K is bounded. When $\mathcal{C}$ is rich enough, it is conceivable that there exists an 'empirical' MV-predictor with mass (*i.e.* coverage probability) $k/(nh^d)$, where k fulfills condition $|k/(nh^d) - \alpha| < c/(nh^d)$. This implies (3.4) since $F(\mathrm{M}_{\mathcal{C}}(\alpha|x)|x) = \alpha$.

Theorem 3.4 (Asymptotic normality of coverage probabilities)
Let conditions (A1) – (A6) *hold, and* (R_γ) *hold with* $\gamma < 1/3$. *Let* $h = c\,n^{-\frac{1}{d+4}}(\log\log n)^{-1}$. *Suppose that* $\mu_{\mathcal{C}}(\cdot|x)$ *is continuous at* α, $F(\mathrm{M}_{\mathcal{C}}(\alpha|x)|x) = \alpha$, *and* (3.4) *holds. Then for* $\tilde{\alpha}_{\mathcal{C}}$ *defined in* (2.7), *we have that as* $n \to \infty$,

$$\sqrt{nh^d}\ (\tilde{\alpha}_{\mathcal{C}} - \alpha) \xrightarrow{d} \mathcal{N}(0,\ \alpha(1-\alpha)\int K^2/f(x)).$$

Remark 3.5 (a) The MV-predictor is not always unique. The above theorem (and also the theorem below) holds for any such a predictor. The existence of an MV-predictor is entailed by the condition (R_γ).???

(b) The only unknown quantity in the asymptotic variance of $\tilde{\alpha}_{\mathcal{C}}$ is the marginal density $f(x)$, which can be estimated consistently. This is in marked contrast with the asymptotic variance of $\widehat{F}(y|x)$. (See, for example, Hall, Wolff and Yao 1999.) Therefore, the confidence level for the coverage probability can be easily constructed based on the above theorem.

(c) The continuity assumption for $\mu_{\mathcal{C}}(\cdot|x)$ is satisfied for all $\alpha \in (0,1)$ if $g(\cdot|x)$ is continuous and has no flat parts, *i.e.* $Leb\{y : g(y|x) = \lambda\} = 0 \ \forall\, \lambda > 0$.

Theorem 3.6 (Asymptotic normality of volumes)
Do you need conditions (A1) – (A6) and etc here? — QY *Suppose that the function* $\mu_{\mathcal{C}}(\cdot|x)$ *defined in* (2.5) *is differentiable and let* $\mu'_{\mathcal{C}}(\cdot|x)$ *denote its Lipschitz continuous derivative. If* $\mu'_{\mathcal{C}}(\alpha|x) > 0$, *then under the assumptions of* Theorem 3.4, *as* $n \to \infty$

$$\sqrt{nh^d}\,\frac{f(x)}{\mu'_{\mathcal{C}}(\alpha|x)}\Big(\widehat{\mu}_{\mathcal{C}}(\alpha|x) - \mu_{\mathcal{C}}(\alpha|x)\Big) \xrightarrow{d} \mathcal{N}(0,\ \alpha(1-\alpha)\int K^2).$$

4 Numerical properties

To appreciate the finite sample properties of the estimated MV-predictors, we illustrate the methods via one nonlinear AR(2) model and a set of the rainfall and river flow data from a catchment in Wales. We always use the standard Gaussian kernel in calculation. We always set the coverage probability $\alpha = 0.9$. We calculate estimators for the MV-interval $M_1(\alpha|x)$ and MV-predictors with at most two intervals $M_1(\alpha|x)$, which will be searched using exhausting method among all the intervals, or unions of two intervals with coverage probability 0.9. We only consider here the estimation of MV-predictors. Examples for estimating predictive distributions can be found in Hall, Wolff and Yao (1999).

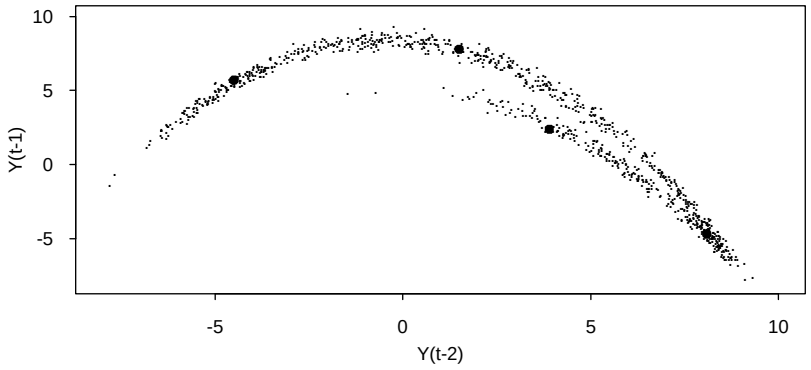

Figure 2: *The scatter plot of* Y_{t-1} *against* Y_{t-2} *from a sample of size 1000 generated from model (4.1). The positions marked with '•' are (from left to right) (-4.5, 5.7), (1.5, 7.8), (3.9, 2.4) and (8.1, -4.7).*

Example 1. We first consider the following simulated model

$$Y_t = 6.8 - 0.17Y_{t-1}^2 + 0.26Y_{t-2} + 0.3\epsilon_t, \tag{4.1}$$

where $\{\epsilon_t\}$ is a sequence of independent random variables each with the standard normal distribution truncated in the interval $[-12, 12]$. We conduct the simulation in two stages to estimate the MV-predictors for Y_t given (i) $X_t \equiv (Y_{t-1}, Y_{t-2})$ and (ii) $X_t \equiv Y_{t-1}$ respectively.

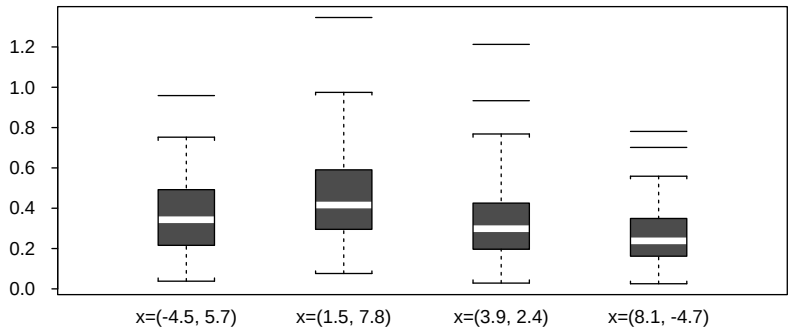

Figure 3: *The boxplots of the ratio of* $\text{Leb}\{\widehat{M}_1(\alpha|x)\Delta M_1(\alpha|x)\}$ *to* $\text{Leb}\{M_1(\alpha|x)\}$ *for model (4.1).*

(i) For four fixed values of $X_t = (Y_{t-1}, Y_{t-2})$, we repeat simulation 100 times with sample size $n = 1000$. Fig.2 is the scatter plot of a sample of size $n = 1000$. The four positions marked with '•' are the values of $X_t = x$ at which the MV-predictors $M_1(\alpha|x)$ for Y_t are estimated. Fig.3 presents the boxplots of $\text{Leb}\{\widehat{M}_1(\alpha|x)\Delta M_1(\alpha|x)\}/\text{Leb}\{M_1(\alpha|x)\}$. The bandwidths were selected by the bootstrap scheme stated in §2.4 based on parametric models determined by AIC. With the given sample sizes, AIC always identified the correct model from the candidature polynomial model of order 3.

(ii) For a sample of size 1000, we estimate both predictors $M_1(\alpha|x)$ and $M_2(\alpha|x)$ for Y_t given its first lagged value $Y_{t-1} = x$ only. We let x range over 90% inner samples. We use a post-sample of size 100 to check the performance of the predictors. For estimating bandwidths using the proposed bootstrap scheme, the parametric model selected by the AIC is

$$Y_t = 8.088 - 0.316Y_{t-1} - 0.179Y_{t-1}^2 + 0.003Y_{t-1}^3 + 0.825\epsilon_t.$$

Fig.4(a) displays the estimated $M_1(\alpha|x)$ together with the 100 post-points. Within the range of values of x on which estimation is conducted, $\widehat{M}_1(\alpha|x)$ contains about 90% of the post-sample. Note that $\widehat{M}_1(\alpha|x)$ has an abrupt change in the width around $x = 1.5$. In fact the predictor

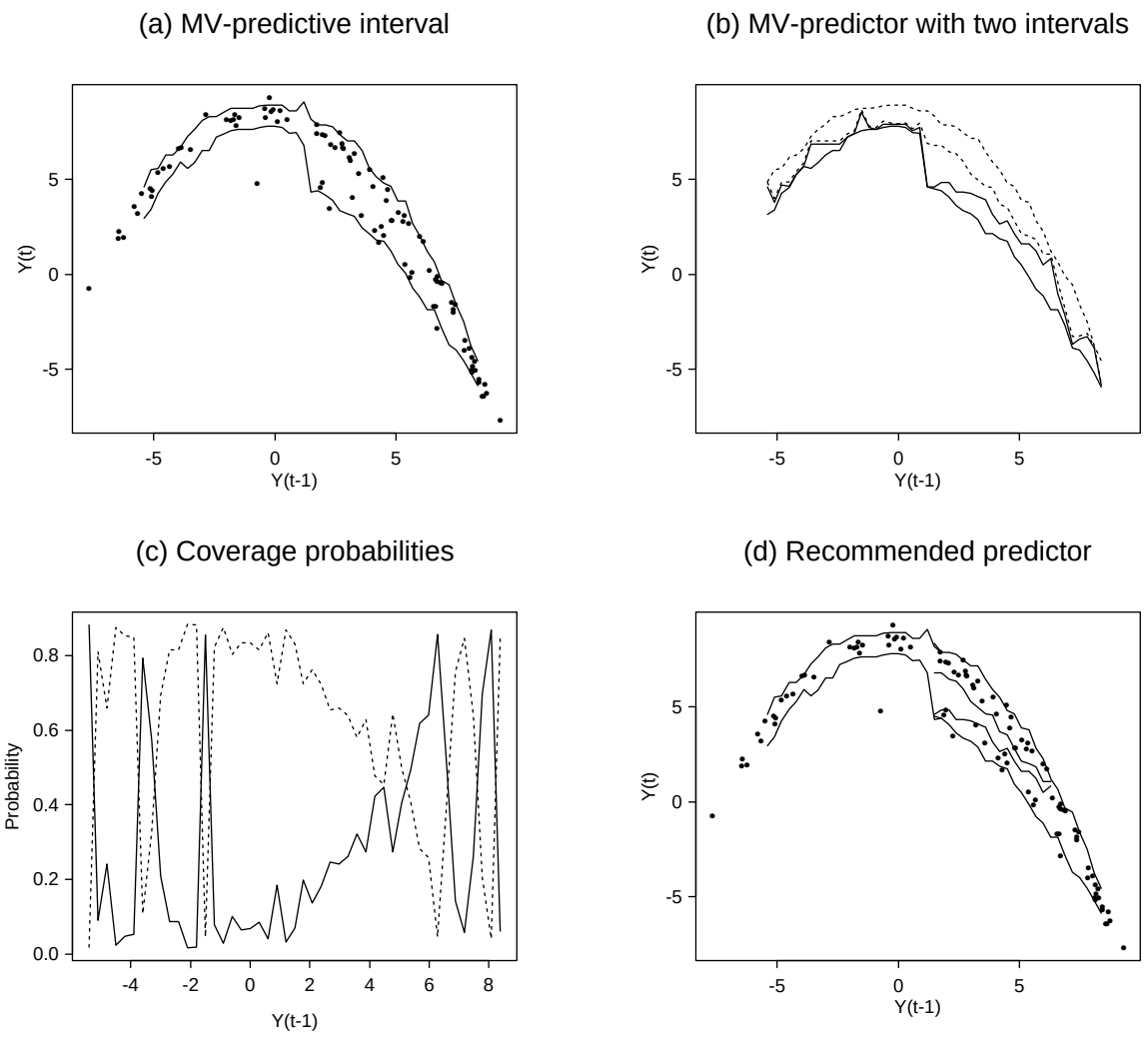

Figure 4: *Simulation results for model (4.1). (a) The estimated* $M_1(\alpha|x)$ *for* Y_t *given* $Y_{t-1} = x$*, together with 100 post-samples. (b) The estimated* $M_2(\alpha|x)$ *for* Y_t *given* $Y_{t-1} = x$*. The two disconnected intervals are bounded respectively with solid lines and dashed lines. (c) The coverage probabilities of the two intervals in (b). Solid curve – the coverage probability for the interval with solid boundary; dashed curve – the coverage probability for the interval with dashed boundary. (d) The recommended MV-predictor for* Y_t *given* $Y_{t-1} = x$*, together with 100 post-samples.*

$M_1(\alpha|x)$ is not satisfactory for x between 1.5 and about 6 because the center of the intervals is void for those x-values (also see Fig.2). The estimator $\widehat{M}_2(\alpha|x)$ is plotted in Fig.4(b). Due to sampling fluctuation, the estimator always consists of two disconnected intervals over the whole sample space. The coverage probabilities of the two intervals are plotted in Fig.4(c). From Fig.4(b) and Fig.4(c), we note that when $x \notin [1.5, 6.3]$, the two intervals of $\widehat{M}_2(\alpha|x)$ are almost connected, and the two corresponding coverage probabilities are either erratic or very close to 0 and 0.9 respectively. Therefore, it seems plausible to use $\widehat{M}_2(\alpha|x)$ for $x \in [1.5, 6.3]$ and $\widehat{M}_1(\alpha|x)$ for $x \notin [1.5, 6.3]$. The combined MV-predictor is depicted in Fig.4(d) together with the post-sample. The combined predictor covers the post-sample as good as the $\widehat{M}_1(\alpha|x)$ (Fig.4(b)) although its Lebesgue measure has be reduced significantly for $x \in [1.5, 6.3]$.

Example 2. Fig.5(a) and Fig.5(b) are plots of 401 hourly rainfall and river flow data from a catchment in Wales. We try to predict the flow from its past values and the rainfall data. Note that the flow data themselves are strongly auto-correlated (Fig.5(c)). Fig.5(d) – (f) indicate that the point-cloud in the scatter plot of flow against rainfall with time lag 2 is slightly thinner than those with time lag 0 and 1, which seems to suggest that the major effect of rainfall on the river flow is of a two-hour delay in time. This is further supported by various statistical modeling procedures. In fact, the cross validation method (Yao and Tong 1994b) specified that the optimal regression subset with two regressors for the flow at the t-th hour Y_t consists of its lagged value Y_{t-1} and the rainfall within the $(t-2)$-th hour X_{t-2}. This was further echoed by a fitted MARS model (Friedman 1991). We now predict Y_t from Y_{t-1} and X_{t-2} using three different types of predictive regions. We estimate the predictors using the data with sample size $n = 394$, which was resulted by leaving out the 373-th, the 375-th and the last five flow data (therefore also their corresponding lagged values and the rainfall data) in order to check the reliability of the prediction. We standardize the observations of regressors before the fitting. We adopt the bootstrap scheme to select the bandwidth. The parametric model determined by AIC is

$$Y_t = -1.509 + 1.191Y_{t-1} + 0.924X_{t-2} + 0.102Y_{t-1}X_{t-2} - 0.004Y_{t-1}^2 + 7.902\epsilon_t,$$

where ϵ_t is standard normal. Table 2 reports the estimated predictors for the 7 data points which are not used in estimation. All the quantile intervals cover the corresponding true values. For the MV-predictor $M_1(\alpha|x)$, 6 out of 7 intervals contain the true value. The only exception occurs when there is high burst of river flow at the value 86.9. It is easy to see from Fig.5 that data

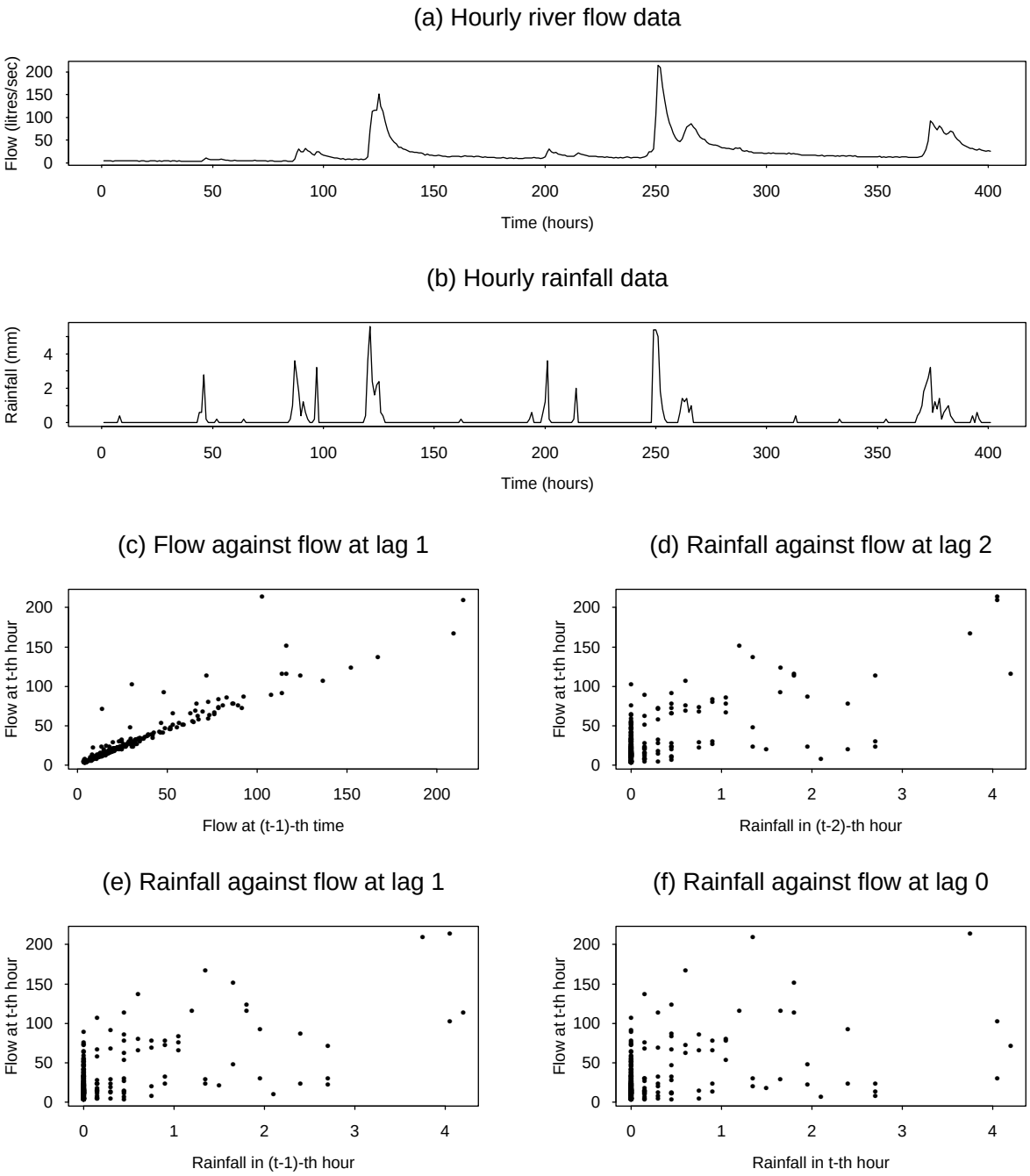

Figure 5: *Hourly rainfall and river flow data from a catchment in Wales. (a) flow (litres/sec); (b) rainfall (mm); (c) scatter plot of flow data; (d) – (f) scatter plots of rainfall against flow at time lags 2, 1 and 0 respectively.*

Table 2. The estimators for the quantile interval $I(\alpha|x)$ and the MV-predictors $M_i(\alpha|x)$ $(i = 1, 2)$ with coverage probability $\alpha = 0.9$ for the river flow at the t-th hour Y_t from its lagged value Y_{t-1} and the rainfall in the $(t$-2)-th hour X_{t-2}.

Y_t	(Y_{t-1}, X_{t-2})	$\widehat{I}(\alpha\|x)$	$\widehat{M_1}(\alpha\|x)$	$\widehat{M_2}(\alpha\|x)$	Coverage probabilities
47.9	(29.1, 1.8)	[3.8, 71.8]	[3.3, 51.1]	[3.6, 42.3]∪[65.4, 67.5]	(0.89, 0.01)
86.9	(92.4, 2.6)	[3.8, 102.9]	[3.3, 76.2]	[3.3, 53.1]∪[62.6, 78.5]	(0.83, 0.07)
28.0	(30.7, 0.6)	[5.6, 39.3]	[3.8, 33.5]	[4.3, 5.6]∪[6.6, 34.6]	(0.03, 0.87)
27.5	(28.0, 0.2)	[4.7, 35.8]	[3.8, 32.4]	[4.1, 26.9]∪[30.7, 34.6]	(0.82, 0.08)
25.4	(27.5, 0.0)	[4.4, 34.6]	[3.8, 32.4]	[3.6, 24.7]∪[30.7, 34.6]	(0.02, 0.88)
26.9	(25.4, 0.0)	[7.7, 33.5]	[9.3, 33.5]	[9.7, 26.9]∪[29.1, 34.1]	(0.81, 0.09)
25.4	(26.9, 0.0)	[4.7, 34.1]	[3.6, 31.7]	[3.3, 25.9]∪[30.7, 34.1]	(0.83, 0.07)

are sparse at this level and upwards. Due to the quick river flow caused by rainfall, we expect the predictive distributions are skewed to the right. Therefore, the improvement in prediction should be observed by using the MV-predictor $M_1(\alpha|x)$ instead of the quantile interval $I(\alpha|x)$. In fact even if we discard the case where the true Y_t lies outside of $\widehat{M_1}(\alpha|x)$, the relative decrease in length of $\widehat{M_1}(\alpha|x)$ with respect to the quantile predictor $\widehat{I}(\alpha|x)$ is between 4.4% and 27.9% for the 6 other cases. Actually $\widehat{M_1}(\alpha|x)$ could be regarded as a compressed shift of $\widehat{I}(\alpha|x)$ to its left in 6 out of 7 cases. For the application with data sets like this, it is pertinent to use the state-dependent bandwidths. For example for estimating $M_1(\alpha|x)$, our bootstrap scheme selected quite large bandwidths 1.57 and 2.99 for the first two cases respectively in Table 2, in response to the sparseness of data in the area with positive rainfall and quick river flow. The selected bandwidths for the last five cases are rather stable and are between 0.33 and 0.43. There seems little evidence suggesting the multi-modality, for the estimated $M_2(\alpha|x)'$s always have one interval with very tiny coverage probability. For this example, we recommend to use the MV-predictive interval $M_1(\alpha|x)$.

In the above application, we include a single rainfall point X_{t-2} in the model for the sake of simplicity. A more pertinent approach should take into account the moisture condition of soil which depends on prior rainfall. For more detailed discussion in this aspect, we refer to Young (1993) and Young and Bevan (1994).

5 Appendix: Proofs

The proofs rely on some asymptotic results for a conditional empirical process which will be published elsewhere. We listed the main results in the Addendum below for easy reference. The

basic idea is similar to the proofs of Polonik (1997) where a global (unconditional) empirical process with *i.i.d.* observations is concerned. The proof for Theorem 3.1 is omitted since it is the least technically involved.

Proof of Theorem 3.2: The basic inequality (5.1) follows along the same lines as (7.24) of Polonik (1997), given in the proof of Theorem 3.1. (We do not present the proof here.) In (5.1) we heavily use the fact that $F(\Gamma_{g(\cdot|x)}(\lambda_\alpha)\Delta \mathrm{M}_{\mathcal{C}}(\alpha)|x) = 0$ which follows from the assumption that $g(\cdot|x) \in \mathcal{M}_{\mathcal{C}}(\alpha)$. For each $\epsilon > 0$ and $M \geq \sup_y g(y|x)$ we have

$$d_{F(\cdot|x)}(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x), \Gamma_{g(\cdot|x)}(\lambda_\alpha)) \;\leq\; D(\epsilon) \;+\; R_{1n}(\epsilon) \;+\; R_{2n}(\epsilon) \;+\; o_P(\frac{1}{\sqrt{nh^d}}), \tag{5.1}$$

$$\begin{aligned}
\text{where} \quad D(\epsilon) &= F(\{y : |g(y|x) - \lambda_\alpha| \leq \epsilon\}|x) \\
R_{1n}(\epsilon) &= \frac{M}{\epsilon}\Big((\widehat{F}_n - F)(\Gamma_{g(\cdot|x)}(\lambda_\alpha)|x) + \frac{1}{\mu'_{\mathcal{C}}(\alpha|x)}(\widehat{\mu}_{\mathcal{C}}(\alpha|x) - \mu_{\mathcal{C}}(\alpha|x))\Big) \\
R_{2n}(\epsilon) &= \frac{M}{\epsilon}\Big((\widehat{F}_n - F)(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha)|x) - (\widehat{F}_n - F)(\Gamma_{g(\cdot|x)}(\lambda_\alpha)|x)\Big).
\end{aligned}$$

Note that in the present situation we have $\frac{1}{\mu'_k(\alpha|x)} = \lambda_\alpha$. Below we use the Bahadur-Kiefer approximation of Theorem 6.2 to control $R_{1n}(\epsilon)$. In fact, it follows that $R_{1n}(\epsilon)$ and $R_{2n}(\epsilon)$ are of the same asymptotic stochastic order.

We now present the proof for $\gamma = 0$. The other cases can be proven similarly, as will be indicated at the end of the proof. First note that

$$D(\epsilon) = O(\epsilon) \tag{5.2}$$

$$R_{1n}(\epsilon) = R_{2n}(\epsilon) = O_P\left(\tfrac{1}{\epsilon\sqrt{nh^d}}\right). \tag{5.3}$$

(5.2) immediately follows from the assumptions. As for (5.3) we use Theorem 6.1 with $\sigma^2 = 1$. By choosing $\epsilon = (nh^d)^{-\frac{1}{4}}$ to balance the rates of the deterministic term $D(\epsilon)$ and the sum of the stochastic terms $R_{1n}(\epsilon) + R_{2n}(\epsilon)$ we get from (5.1) a first stochastic rate

$$d_{F(\cdot|x)}(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x), \Gamma_{g(\cdot|x)}(\lambda_\alpha)) = O_P\left((nh^d)^{-\frac{1}{4}}\right).$$

Note that here at this first step we could choose an even larger bandwidth, namely the optimal bandwidth $h = n^{-\frac{1}{d+4}}$. Then, however, we would have to stop here, and this would finally give us a slower rate of convergence. Instead, by using the faster bandwidth as given in the Theorem, this first rate can be used to obtain a faster rate of convergence (and then go on iterating). By choosing $\sigma^2 = (nh^d)^{-\frac{1}{4}}$ we obtain from Theorem 6.1 and Theorem 6.2, respectively, that

$$R_{1n}(\epsilon) = R_{2n}(\epsilon) = O_P\left(\frac{1}{\epsilon}(nh^d)^{-\left(\frac{1}{2}+\frac{1}{8}\right)}(\log n)^{\frac{1}{2}}\right). \tag{5.4}$$

As above, by balancing the deterministic and the stochastic term in (5.1) we get the second rate

$$d_{F(\cdot|x)}(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x), \Gamma_{g(\cdot|x)}(\lambda_\alpha)) = O_P\left((nh^d)^{-(\frac{1}{4}+\frac{1}{16})}(\log n)^{\frac{1}{4}}\right).$$

Iterating this argument M times we obtain the rate $(nh^d)^{-S(M)}(\log n)^{S(M-1)}$ where $S(M) = \sum_{j=1}^{M}(\frac{1}{4})^j$. We can do this iteration arbitrarily, but finitely, often. Since $S(M) \to 1/3$ as $M \to \infty$ the assertion follows.

Note that in order to assure that the assumptions of Theorem 6.1 on h and σ^2 are satisfied in each iteration step we need for each $\eta > 0$ that $nh^{d+4} \leq (nh^d)^{-\frac{1}{3}+\eta}$. This leads to the condition $h = c\, n^{-\frac{1}{d+3}}$.

As for $\gamma > 0$, the same proof in principle works. However, we have to take into account that we can only proceed the above iteration steps as long as the resulting rate is not faster than $\left(nh^d\right)^{\frac{3\gamma-1}{2(3\gamma+1)}}$, which comes from the definition of Λ_γ in Theorem 6.1. Due to this reason, we can do this iteration arbitrarily often only for $\gamma < 1/5$. For $\gamma \geq 1/5$ iteration does not help, so that the arbitrary $\eta > 0$ in the exponent does not appear.

q.e.d.

Proof of Theorem 3.4: By assumption we have

$$F(\mathrm{M}_{\mathcal{C}}(\alpha|x)|x) = \alpha = \widehat{F}_n(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x)|x) + o_P\left(\frac{1}{\sqrt{nh^d}}\right)$$

it follows that

$$\begin{aligned}\sqrt{nh^d}\ (\tilde{\alpha}_{\mathcal{C}} - \alpha) &= \sqrt{nh^d}\ (F(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x)|x) - \widehat{F}_n(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x)|x)) + o_P(1)\\ &= \sqrt{nh^d}\ (F(\mathrm{M}_{\mathcal{C}}(\alpha|x)|x) - \widehat{F}_n(\mathrm{M}_{\mathcal{C}}(\alpha|x)|x)) + o_P(1).\end{aligned}$$

The last equality follows from Theorem 6.1, because Theorem 3.1 gives consistency of $\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x)$, this is $d_{F(\cdot|x)}(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x), \mathrm{M}_{\mathcal{C}}(\alpha|x)) = o_P(1)$. Finally, asymptotic normality of $\sqrt{nh^d}\ (F(\mathrm{M}_{\mathcal{C}}(\alpha|x)|x) - \widehat{F}_n(\mathrm{M}_{\mathcal{C}}(\alpha|x)|x))$ with given variance under the present conditions is not difficult to proof. It follows by nowadays almost standard arguments. The proof is therefore omitted.

q.e.d.

Theorem 3.6 immediately follows from Theorem 6.2 together with the abovementioned asymptotic normality of the conditional empirical process.

6 Addendum:

For ease of reference we present two theorems about the asymptotic behavior of a conditional empirical process. The proofs of these results will be published elsewhere.

Let

$$\Lambda_\gamma(\sigma^2, n) = \begin{cases} \sqrt{\sigma^2 \log \frac{1}{\sigma^2}} & \text{if } \gamma = 0, \\ \max\left((\sigma^2)^{\frac{1-\gamma}{2}}, \left(nh^d\right)^{\frac{3\gamma-1}{2(3\gamma+1)}} \right) & \text{if } \gamma > 0. \end{cases} \tag{6.1}$$

Theorem 6.1 *Suppose that* (A2) – (A6) *hold. For each* $\sigma^2 > 0$, *let* $\mathcal{C}_\sigma \subset \mathcal{C}$ *be a class of measurable sets with* $\sup_{C \in \mathcal{C}_\sigma} F(C|x) \leq \sigma^2 \leq 1$, *and suppose that* $\mathcal{C}$ *fulfills* (R_γ) *with some* $\gamma \geq 0$. *Further we assume that* $h^d \to 0$ *and* $nh^d \to \infty$ *as* $n \to \infty$ *such that*

$$nh^{d+4} \leq \left(\Lambda_\gamma(\sigma^2, n)\right)^2. \tag{6.2}$$

For $\gamma = 0$ *we assume in addition that* $\frac{nh^d\sigma^2 \log \frac{1}{\sigma^2}}{(\log n)^6} \to \infty$ *as* $n \to \infty$. *Then there exists a constant* $M > 0$ *such that* $\forall\, \epsilon > 0,\ \exists\, \sigma_0^2 > 0$ *such that for all* $\sigma^2 \leq \sigma_0^2$ *and for large enough* n

$$P\left(\sup_{C \in \mathcal{C}_\sigma} |\nu_n(C|x)| \geq M\, \Lambda_\gamma(\sigma^2, n) \right) \leq \epsilon.$$

Theorem 6.2 (Generalized Bahadur-Kiefer approximation)
Suppose that (A2) – (A6) *hold. Assume that* $\mu_{\mathcal{C}}(\cdot|x)$ *is differentiable with Lipschitz-continuous derivative* $\mu'_k(\cdot|x)$. *Let* $\mathcal{C}$ *be such that* (R_γ) *holds. Let further* $\alpha \in (0,1)$ *be fixed and suppose that* $\mathrm{M}_{\mathcal{C}}(\alpha|x)$ *is unique up to Leb-nullsets, that* $F(\mathrm{M}_{\mathcal{C}}(\beta|x)|x) = \beta$ *for all* β *in a neighborhood of* α, *and that* $\mu'_{\mathcal{C}}(\alpha|x) > 0$. *If for* h *and* σ^2 *satisfying the conditions of* Theorem 6.1 *we have for* $n \to \infty$ *that*

$$d_{F(\cdot|x)}(\widehat{\mathrm{M}}_{\mathcal{C}}(\alpha|x), \mathrm{M}_{\mathcal{C}}(\alpha|x)) = O_P(\sigma^2),$$

then for $n \to \infty$

$$\sqrt{nh^d}\ |(\widehat{F}_n - F)(\mathrm{M}_{\mathcal{C}}(\alpha|x)) + \frac{1}{\mu'_{\mathcal{C}}(\alpha|x)}(\widehat{\mu}_{\mathcal{C}}(\alpha|x) - \mu_{\mathcal{C}}(\alpha|x))| = O_P\left(\Lambda_\gamma(\sigma^2, n)\right).$$

References

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rodgers, W.H., and Tukey, J.W. (1972). *Robust estimation of location: survey and advances.* Princeton Univ. Press, Princeton, N.J.

Bosq, D. (1996). *Nonparametric statistics for stochastic processes.* Lecture Notes in Statistics No. 110, Springer, New York.

Brännäs, K. and De Gooijer, J.G. (1992). Modelling business cycle data using autoregressive-asymmetric moving average models. *ASA Proceedings of Business and Economic Statistics Section*, 331- 336.

Chan, K.S. and Tong, H. (1994). A note on noisy chaos. *J. R. Statist. Soc.* **B**, **56**, 301-311.

Davies, N., Pemberton, J. and Petruccelli, J.D. (1988). An automatic procedure for identification, estimation and forecasting univariate self exciting threshold autoregressive models. *The Statistician*, **37**, 119-204.

De Gooijer, J.G. and Gannoun, A. (1997). Nonparametric conditional predictive regions for stochastic processes. (A preprint.)

Dudley, R.M. (1974). Metric entropy of classes of sets with differentiable boundaries. *J. Approx. Theorie* **10** 227-236.

Dudley, R.M. (1984). A course in empirical processes. *École d'Été de Probabilités de Saint Flour XII-1982, Lecture Notes in Math.* **1097** 1-142, Springer, New York.

Einmahl, J.H.J. and Mason, D.M. (1992). Generalized quantile processes. *Ann. Statist.* **20**, 1062-1078.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.*, **87**, 998-1004.

Fan, J., Hu, T.C. and Truong, Y.K. (1994). Robust nonparametric function estimation. *Scand. J. Statist.* **21**, 433-446.

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

Friedman, J.H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19**, 1-142.

Hall, P., Wolff, R.C.L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *J. Ameri. Statist. Assoc.* **94**, 154-163.

Hartigan, J.A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82** 267-270

Hyndman, R.J. (1995). Highest density forecast regions for non-linear and non-normal time series models. *J. Forecasting*, **14**, 431-441.

Hyndman, R.J. (1996). Computing and graphing highest density regions. *Ameri. Statist.* **50**, 120-126.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Stat.* **18**, 191-219 Paris-South, Orsay

Müller, D.W. and Sawitzki, G. (1987). Using excess mass estimates to investigate the modality of a distribution. Preprint No. 398, SFB 123, Universität Heidelberg

Müller, D.W. and Sawitzki, G. (1991). Excess mass estimates and tests of multimodality. *J. Amer. Statist. Assoc.* **86** 738-746

Nolan, D. (1991). The excess mass ellipsoid. *J. Multivariate Anal.* **39** 348 - 371

Polonik, W. (1995). Measuring mass concentration and estimating density contour clusters - an excess mass approach. *Ann. Statist.* **23** 855-881

Polonik, W. (1997). Minimum volume sets and generalized quantile processes. *Stoch. Processes and Appl.* **69** 1-24

Polonik, W. and Yao, Q. (1998). Asymptotics of set-indexed conditional empirical processes based on dependent data. Unpublished manuscript.

Silverman, B.W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc.* **B**, **43**, 97-99.

Tsybakov, A. B. (1986). Robust function reconstruction by local approximation. *Problems of Information Transmission* **22** 69 - 84.

Tong, H. and Moeanaddin, R. (1988). On multi-step non-linear least squares prediction. *The Statistician*, **37**, 101-110.

Yao, Q. and Tong, H. (1994a). Quantifying the inference of initial values on nonlinear prediction. *J. Roy. Statist. Soc.* **B**, 56, 701-725.

Yao, Q. and Tong, H. (1994b). On subset selection in non-parametric stochastic regression. *Statistica Sinica*, **4**, 51-70.

Yao, Q. and Tong, H. (1995a). Asymmetric least squares regression estimation: a nonparametric approach. *J. Nonparametric Statist.*, **6**, 273-292.

Yao, Q. and Tong, H. (1995b). On initial-condition sensitivity and prediction in nonlinear stochastic systems. *Bull. Int. Statist. Inst.*, **IP 10.3**, 395-412.

Yu and Jones (1998). Local linear quantile regression. *J. Amer. Statist. Assoc.* **93**, 228-237.

Young, P.C. (1993). Time variable and state dependent modelling of nonstationary and nonlinear time series. *Developments in Time Series Analysis*, ed. T.Subba Rao. Chapman and Hall, London, 374-413.

Young, P.C., and Beven, K.J. (1994). Data-based mechanistic modelling and the rainfall-flow nonlinearity. *Environmetrics*, **5**, 335-363.