Statistica Sinica **26** (2016), 1649-1672
doi:http://dx.doi.org/10.5705/ss.2014.196t

FUNCTIONAL COEFFICIENT MOVING AVERAGE MODEL WITH APPLICATIONS TO FORECASTING CHINESE CPI

Song Xi Chen[1,2], Lihua Lei[1] and Yundong Tu[1]

[1]*Peking University and* [2]*Iowa State University*

Abstract: This article establishes a functional coefficient moving average model (FMA) that allows the coefficient of the classical moving average model to adapt with a covariate. The functional coefficient is identified as a ratio of two conditional moments. A local linear estimation technique is used for estimation and the asymptotic properties of the resulting estimator are investigated. Its convergence rate depends on whether the underlying function reaches its boundary or not, and the asymptotic distribution can be nonstandard. A model specification test in the spirit of Härdle-Mammen (1993) is developed to check the stability of the functional coefficient. Simulations have been conducted to study the finite sample performance of our proposed estimator, and the size and the power of the test. Application is made to CPI data from the China Mainland and to German egg prices to show the efficacy of FMA.

Key words and phrases: Consumer price index, forecasting, functional coefficient model, moving average model.

1. Introduction

Autoregressive Integrated Moving Average (ARIMA) models have been popular in time series analysis due to their simplicity and adaptability. An ARIMA (p, d, q) model can be expressed as:

$$(1-B)^d(1-\phi_1 B-\cdots-\phi_p B^p)x_t = \mu + (1+\theta_1 B+\cdots+\theta_q B^q)\epsilon_t$$

where B is the lagged operator and $\{\epsilon_t\}$ is a white noise series with zero mean and finite variance. This describes a special dependence structure of the data that can be regarded as an approximation to all stationary process according to the Wold Decomposition Theorem. In the past decades, numerous works in statistics and econometrics have been devoted to studying and extending the ARIMA model and its applications (for example, Box and Jenkins (1976); Box and Tiao (1975); Dahlhaus (1989); Cleveland and Tiao (1976); Granger and Joyeux (1980); Hannan and Deistler (1988)).

One important application of ARIMA model is to forecast the Consumer Price Index (CPI). The growth rate of CPI can be regarded as a proxy for the

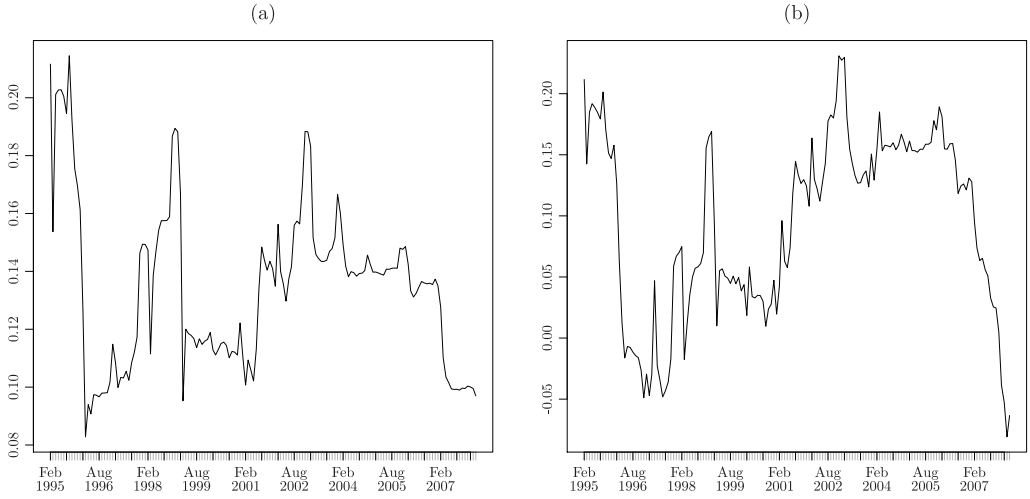

Figure 1. Estimates of θ: (a) expanding window; (b) rolling window.

inflation rate, which is a chief target of macro-economic management by various governments and is an important economic indicator for investors. One popular model for the CPI is ARIMA(0,1,1) (Nelson and Schwert (1977); Schwert (1987); Barsky (1987)):

$$(1 - B)x_t = \mu + (1 - \theta B)\epsilon_t,$$

where x_t represents logarithm of CPI. Although the model is easy to implement, it restricts the inflation dynamics to have an autocovariance that is constant over time. Meanwhile, for the US data, the estimates of θ are not stable over time and indeed are fairly volatile. Stock and Watson (2006) interpreted this instability as the variation of variance, which changes inversely with the magnitude of MA coefficient estimates. Parameter instability is also observed in our analysis when analyzing monthly CPI data of the China Mainland from January 1990 to March 2014. We build an ARIMA(0,1,1) model on the year-on-year CPI monthly growth data, and estimate the MA coefficient θ on an expanding window basis and a rolling window basis with a 60-month window-width. These estimates are plotted in Figure 1. It can be seen that the estimates of θ are quite variable.

Accordingly, we consider an extension of ARIMA(0,1,1) model in which the MA coefficient is a smooth function of a state covariate z_t such that

$$(1 - B)x_t = \mu + (1 - \theta(z_t)B)\epsilon_t. \tag{1.1}$$

This is called the Functional Moving Average (FMA) model of order 1, or FMA(1). The state variable z_t contains information that affects the dynamics of x_t, and does not have to be exogeneous. The dynamics available to z_t

are very general, as indicated by the Assumptions (A3)−(A5) given in Section 2.2. The choice of z_t can be made based on, for example, related economic theory, or through a data driven procedure. We provide a testing procedure that determines whether a given variable is qualified as a state variable that can be used to improve the inference and prediction of x_t, in Section 2.4. We focus on inference on FMA(1); extension to higher order FMA will be discussed in the conclusion. Our FMA model is related to the state-dependent models of Priestley (1980) and to the autoregressive functional moving average (ARFMA) model of Wang (2008), where the latter is a specific form of the former. However, the ARFMA model has the functional coefficient of the MA parts being functions of the lagged values of the variable x_t itself, while ours need not. The two can be united under a more general framework with multivariate state variables. In any case, the asymptotic properties of the estimators for the state-dependent and the ARFMA models are not known. Here we provide them for the FMA(1) model.

In econometrics and time series literatures (Hamilton (1994)), the MA coefficients are often explained as the Impulse Response (IR). Thus, for any series x_t that can be written in the MA(∞) form

$$x_t = \mu + \sum_{j\geq 0} \theta_j \epsilon_{t-j},$$

the jth order IR is $\partial x_t/\partial \epsilon_{t-j} = \theta_j$ for any $j \geq 0$. This measures the effect of a shock on the response after j periods. For the FMA(1) model, the 1-st order IR is $\theta(z_t)$, a function of the state variable rather than a constant as in the MA(1) model. This flexibility brings closer linkage to realism, as the effect of a shock is often affected by the state of the world.

Our work is closely related to a large body of literature on varying coefficient models. They have been well developed in nonparametric statistics and time series analysis, including ARCH/GARCH (Engle (1982); Bollerslev (1986)), TAR (Tong (1983); Chan and Tong (1986); Tong (1990); Tiao and Tsay (1994); Caner and Hansen (2001), EXPAR (Haggan and Ozaki (1981); Ozaki (1982)) and FAR (Chen and Tsay (1993); Fan, Yao, and Cai (2003)). This literature focuses mainly on extending the AR component of the ARIMA model, while the current work aims to relax the flexibility of the MA component. See also Priestley (1980) and Wang (2008).

The unique feature in the inference for the FMA(1) model is the estimation technique. Unlike the FAR(1) model which has a regression form, local polynomial regression cannot be directly applied to FMA (1). Nevertheless, we find that the functional coefficient is identified via the conditional autocovariance function. As a result, the functional coefficient can be consistently estimated by first estimating the autocovariance function. To this end, local linear least square is used to obtain estimates of conditional moments.

This paper can be extended in several directions. An AR component can be incorporated to allow for more general dependence structure, and the FMA(1) model can be generalized to allow for multiple state variables Z_t. To avoid the curse of dimensionality, a single index structure for $\theta(\cdot)$, such as $\theta(Z_t^\top \gamma)$, could be imposed and the estimation procedure adapted from Ichimura (1993) used. The resulting identification and estimation problem is more involved and deserves a separate treatment. We leave these extensions for future research.

The rest of paper is structured as follows. The next section introduces the details for identification and estimation of the FMA model. The asymptotic distribution of the proposed estimator is established, and a model specification test is developed. Section 3 presents simulation results that evaluate the finite sample performance of our estimator, and the size and power of the model specification test. Section 4 shows the efficacy of FMA model by forecasting Chinese CPI data and comparing this to MA(1) models. Section 5 concludes with remarks on future work. Technical lemmas and all proofs are in the online supplementary materials.

2. Theoretical Property

2.1. Identification and estimation

For the MA(1) model

$$x_t = \mu + \epsilon_t + \theta\epsilon_{t-1},$$

where $\{\epsilon_t\}$ is a white noise process with variance σ^2, the variance and the first autocovariance of x_t are

$$\begin{aligned} E((x_t-\mu)^2) &= (1+\theta^2)\sigma^2, \\ E((x_t-\mu)(x_{t-1}-\mu)) &= \theta\sigma^2. \end{aligned}$$

Higher order autocovariances are all 0's. Then θ can be estimated via the ratio of two moments after certain transformation.

Suppose x_t follows an FMA model with the state variable z_t,

$$x_t = \mu + \epsilon_t + \theta(z_t)\epsilon_{t-1},$$

where $\{\epsilon_t\}$ is a white noise with variance σ^2, $\theta(z_t)$ is a smooth function with $|\theta(z_t)| \le 1$. Here,

$$\begin{aligned} E((x_t-\mu)^2|z_t=z) &= E(\epsilon_t^2|z_t=z) + 2\theta(z)E(\epsilon_t\epsilon_{t-1}|z_t=z) + \theta^2(z)E(\epsilon_{t-1}^2|z_t=z), \\ E((x_t-\mu)(x_{t-1}-\mu)|z_t=z) &= E(\epsilon_t\epsilon_{t-1}|z_t=z) + E(\theta(z_{t-1})\epsilon_t\epsilon_{t-2}|z_t=z) \\ &\quad +\theta(z)\{E(\epsilon_{t-1}^2|z_t=z) + E(\theta(z_{t-1})\epsilon_{t-1}\epsilon_{t-2}|z_t=z)\}. \end{aligned}$$

If for $j, k = 0, 1$,

$$E(\epsilon_{t-k}\epsilon_{t-j}|z_t) = E(\epsilon_{t-k}\epsilon_{t-j}) = \sigma^2 I(j = k) \text{ and} \tag{2.1}$$

$$E(\epsilon_{t-j}\epsilon_{t-2}|z_t, z_{t-1}) = E(\epsilon_{t-j}\epsilon_{t-2}) = 0, \tag{2.2}$$

then

$$E((x_t - \mu)^2|z_t = z) = (1 + \theta^2(z))\sigma^2 \text{ and} \tag{2.3}$$

$$E((x_t - \mu)(x_{t-1} - \mu)|z_t = z) = \theta(z)\sigma^2. \tag{2.4}$$

The two conditional moments have the same form as those of the MA(1) model. The condition (2.1) and (2.2) are satisfied if (z_t, z_{t-1}) is independent of $(\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2})$ for all t. In practice, z_t is often taken as lagged variables that contain the state information, as in FAR (Cai, Fan, and Yao (2000)). This requirement is not as stringent as it appears, since it is often reasonable to assume the independence between the future innovations and the past observations. This condition is precisely described in Assumption (A5) below.

Nonparametric method of moments can be used to estimate $\theta(z)$. To do so, we need to estimate the conditional moments in (2.3) and (2.4). Nonparametric estimators such as the Nadaraya-Watson estimator (Nadaraya (1964); Watson (1964)) and the local polynomial estimator (Fan and Gijbels (1996)) could be used. We prefer the local linear estimators due to such properties as minimax efficiency, automatic boundary correction and a simpler form of the asymptotic bias. Denote the local linear estimator of the variance and the autocovariance by $\hat{a}_0(z)$ and $\hat{a}_1(z)$, and solve

$$(\hat{a}_j(z), \hat{b}_j(z)) = \underset{(a,b)}{\operatorname{argmin}} \sum_{t=1}^{T} \{(x_t - \bar{x})(x_{t-j} - \bar{x}) - a - b(z_t - z)\}^2 K(\frac{z_t - z}{h}),$$

for $j = 0, 1$, where $\bar{x} = T^{-1}\sum_{t=1}^{T} x_t$ is a consistent estimator for μ, $k(\cdot)$ is a kernel function, and h is the smoothing parameter.

Let $g(w) = w/(1 + w^2)$, monotone in $w \in [-1, 1]$. A natural estimator for $g\{\theta(z)\}$ is

$$\hat{g}\{\theta(z)\} = \frac{\hat{a}_1(z)}{\hat{a}_0(z)}. \tag{2.5}$$

Here $|g(w)| \leq 1/2$ for all $w \in [-1, 1]$. To incorporate this restriction, we consider the constrained estimator

$$\tilde{g}\{\theta(z)\} = \hat{g}\{\theta(z)\}I(|\hat{g}\{\theta(z)\}| \leq \tfrac{1}{2}) + \tfrac{1}{2}I(\hat{g}\{\theta(z)\} > \tfrac{1}{2}) - \tfrac{1}{2}I(\hat{g}\{\theta(z)\} < -\tfrac{1}{2}). \tag{2.6}$$

Then, $\theta(z)$ can be estimated by $\hat{\theta}(z) = h(\tilde{g}\{\theta(z)\})$, where $h : [-1/2, 1/2] \to [-1, 1]$, and

$$h(x) = g^{-1}(x) = \begin{cases} \frac{1-\sqrt{1-4x^2}}{2x} & (\text{ if } x \neq 0); \\ 0 & (\text{ if } x = 0). \end{cases}$$

2.2. Large sample theory

To ease the presentation, we only consider the case where z_t is a scalar. The extension to multi-dimensional state variables follows in a similar fashion. We need some regularity conditions.

(A1) $h = O(T^{\epsilon_0 - 1})$ as $T \to \infty$ for some $\epsilon_0 \in (0, 1)$.

(A2) The kernel function $K(\cdot)$ is symmetric, has support $S_K = [-1, 1]$, and there exists a $M > 0$ such that $|K(x) - K(y)| \leq M|x - y|$ for all $x, y \in S_K$.

(A3) (i) $\{\epsilon_t\}$ is a white noise sequence with $E\epsilon_t^2 = \sigma^2 < \infty$, $E|\epsilon_t|^{2\delta} < \infty$ for some $\delta > 2$; (ii) $\{\epsilon_t, z_t\}$ is a strictly stationary α-mixing process with the mixing coefficients satisfying the condition $\alpha(k) < ck^{-\beta}$ for some $\beta > \max\{(2\delta - 2)/(\delta - 2), (2 - \epsilon_0)/\epsilon_0\}$ and constant $c > 0$.

(A4) (i) The density function $p(z)$ of z_t has a bounded second derivative; (ii) the conditional density function of (z_1, z_m) given $(x_1, \ldots, x_m)$ is bounded by a constant C_0 uniformly with $m \geq 0$; (iii) the conditional density of x_t given z_t is continuous.

(A5) (i) For each t and $j, k = 0, 1$, $E(\epsilon_{t-k}\epsilon_{t-j}|z_t) = \sigma^2 I(j = k)$ and $E(\epsilon_{t-j}\epsilon_{t-2}|z_t, z_{t-1}) = 0$; (ii) $E(|\epsilon_{t-j}|^{2\delta}|z_t = z) \leq M < \infty$ for some M and $j = 0, 1, 2$, with δ as in (A2).

(A6) The coefficient function $\theta(z)$ has a continuous second derivative and $|\theta(z)| \leq 1$ for any $z \in \mathbb{R}$.

Conditions (A1) and (A2) are standard and we use the second-order Epanechnikov kernel throughout. Conditions (A3) and (A4) have been used by Masry and Fan (1997) for α-mixing processes. The condition imposed on β in (A3) is a technical requirement. If $Ee^{\lambda|\epsilon_t|^\alpha} < \infty$ for some $\lambda, \alpha > 0$, then δ can be arbitrarily large and hence (A3) can be reduced to $\beta > 2$ if $\epsilon_0 > 2/3$. Condition (A5.i) is needed for identification of the model, and (A5.ii) is a technical condition in order to apply the result of Masry and Fan (1997); it holds under (A3) if z_t is independent of $(\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2})$. (A6) places smoothness condition on the functional coefficient. In particular, z_t does not have to be exogeneous. The dynamics available to z_t are very general as indicated by (A3)$-$(A5), which are largely for the mixing condition, the conditional moment conditions and conditions regarding the conditional densities.

We begin with the asymptotic normality of $\hat{g}\{\theta(z)\}$. Let

$$G(z) = \frac{u(z)^\top A u''(z)}{2[1 + \theta^2(z)]^2}\sigma_K^2, \nu(z) = \frac{u(z)^\top \Gamma(z) u(z)}{[1 + \theta^2(z)]^4 p(z)} R(K), A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \text{ and}$$

$$\Gamma(z) = \frac{Cov\left[(x_t-\mu)(x_{t-1}-\mu),(x_t-\mu)^2|z_t=z\right]}{\sigma^4},$$

where $\sigma_K^2 = \int u^2K(u)du$, $R(K) = \int K^2(u)du$, $u(z) = (1+\theta^2(z), -\theta(z))^\top$, and $\theta'(z)$ and $\theta''(z)$ are the first and second derivatives of $\theta(z)$. Let $\mathbb{S} = \{z|p(z) > 0\}$.

Theorem 1. *Under* (A1)∼(A6)*, for* $z \in \mathbb{S}$*, as* $T \to \infty$*,*

$$\sqrt{Th}(\hat{g}\{\theta(z)\} - g\{\theta(z)\} - G(z)h^2) \xrightarrow{d} N(0, \nu(z)).$$

Remark 1. Let $\mathbb{M} = \{z : \theta(z) = \pm 1\}$. Since $|\theta(z)| \le 1$ for all $z \in \mathbb{R}$, the points in $\mathbb{M}$ are local extrema of $\theta(z)$. Thus, $\theta'(z) = 0$ for all $z \in \mathbb{M}$. It is easily shown that $G(z) = 0$, for $z \in \mathbb{M}$.

Theorem 2. *Under* (A1)∼(A6)*, for* $z \in \mathbb{S}$*,*

(i) *If* $|g\{\theta(z)\}| < 1/2$, $\sqrt{Th/\nu(z)}(\tilde{g}\{\theta(z)\} - g\{\theta(z)\} - G(z)h^2) \xrightarrow{d} \Phi$;

(ii) *If* $g\{\theta(z)\} = 1/2$, $\sqrt{Th/\nu(z)}(\tilde{g}\{\theta(z)\} - g\{\theta(z)\}) \xrightarrow{d} \Phi^-$;

(iii) *If* $g\{\theta(z)\} = -1/2$, $\sqrt{Th/\nu(z)}(\tilde{g}\{\theta(z)\} - g\{\theta(z)\}) \xrightarrow{d} \Phi^+$,

where Φ *is the standard normal distribution function, and*

$$\Phi^-(x) = \Phi(x)I(x<0) + I(x \ge 0), \Phi^+(x) = \Phi(x)I(x \ge 0).$$

This result reveals the distribution discontinuity at the boundaries of $g\{\theta(z)\}$. Intuitively, when $|g\{\theta(z)\}| < 1/2$, the unconstrained estimator $\hat{g}\{\theta(z)\}$ is the same as $\tilde{g}\{\theta(z)\}$ for sample size large enough, hence asymptotically equivalent. However, when $|g\{\theta(z)\}| = 1/2$, $\hat{g}\{\theta(z)\} \ne \tilde{g}\{\theta(z)\}$, with positive probability, and the asymptotic distributions differ.

We are in a position to state the asymptotic property of $\hat{\theta}(z) = h(\tilde{g}\{\theta(z)\})$, noting that $h(x)$ is differentiable when $|x| < 1/2$. The delta-method can be applied to Theorem 2 to obtain the asymptotic distribution of $\hat{\theta}(z)$. At $|x| = 1/2$, the asymptotic distribution can be derived directly. See the appendix for details.

Theorem 3. *Under* (A1)∼(A6)*, for* $z \in \mathbb{S}$*,*

(i) *If* $|\theta(z)| < 1$, $\sqrt{Th/\nu(z)}g'\{\theta(z)\}(\hat{\theta}(z) - \theta(z) - g'\{\theta(z)\}^{-1}G(z)h^2) \xrightarrow{d} \Phi$;

(ii) *If* $\theta(z) = 1$, $\sqrt[4]{Th/\nu(z)}(\hat{\theta}(z) - \theta(z)) \xrightarrow{d} H_\Phi^-$;

(iii) *If* $\theta(z) = -1$, $\sqrt[4]{Th/\nu(z)}(\hat{\theta}(z) - \theta(z)) \xrightarrow{d} H_\Phi^+$,

where $H_\Phi^-(x) = \Phi(-x^2/4)I(x<0) + I(x \ge 0)$ *and* $H_\Phi^+(x) = \Phi(x^2/4)I(x \ge 0)$.

When $\theta(z) = \pm 1$, convergence is at a slower rate and the asymptotic distribution is nonstandard. The proposed ratio estimator of $\theta(\cdot)$ function may not be efficient; more efficient estimation is to investigated in a separate endeavour.

The asymptotic variance of the estimators depend on the unknown parameter σ^2. It can be consistently estimated by the sample average of the squared innovation residuals $\hat{\epsilon}_t^2$, for $t = 1, \ldots, T$ under (A3), where the $\hat{\epsilon}_t$ can be obtained in a similar iterative procedure as in moving average models, with $\hat{\theta}$ replaced by the estimated function $\hat{\theta}(z_t)$.

2.3. Bandwidth selection

The theoretical optimal bandwidth for estimating $\theta(z)$ that minimizes the asymptotic mean squared error of $\hat{\theta}(z)$ can be written as

$$\hat{h}^{opt} = (\frac{\nu(z)g'(\theta(z))^2}{4G(z)^2T})^{1/5} = \left(c_K \frac{u(z)^\top \Gamma(z)u(z)g'(\theta(z))^2}{u(z)^\top \Lambda(z)u(z)p(z)}\right)^{1/5} T^{-1/5}, \tag{2.7}$$

where $c_K = R(K)/\sigma_K^4$ and $\Lambda(z) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} u''(z)u''^\top(z) \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. This theoretically optimal bandwidth depends on the unknown elements $\theta(z)$, $\Gamma(z)$, $\Lambda(z)$, and $p(z)$. In practice, these terms can be consistently estimated with a prior bandwidth.

A practical way to bandwidth selection adopts the Residual Squares Criterion (RSC) proposed by Fan and Gijbels (1995), which avoids these complications. Let

$$\hat{\Gamma}(z,h) = \frac{1}{\Delta}\sum_{t=2}^{T}(Y_t - \hat{Y}_t)(Y_t - \hat{Y}_t)^\top K(\frac{z_t - z}{h})$$

where $\Delta = tr(W - WZ(Z'WZ)^{-1}Z'W)$, $Z = [(1, z_2 - z)^\top, \ldots, (1, z_T - z)^\top]^\top$, $W = diag\{K((z_2 - z)/h), \ldots, K((z_T - z)/h)\}$, and $Y_t = ((x_t - \mu)^2, (x_t - \mu)(x_{t-1} - \mu))^\top$; here $\hat{Y}_t = (\hat{a}_0^*(z_t), \hat{a}_1^*(z_t))^\top$, where

$$(\hat{a}_j^*(z), \hat{b}_j^*(z)) = \underset{a,b}{\operatorname{argmin}} \sum_{t=1}^{T}\{(x_t - \mu)(x_{t-j} - \mu) - a - b(z_t - z)\}^2 K(\frac{z_t - z}{h}).$$

With similar arguments as in Fan and Gijbels (1995), it can be shown that

$$E(\hat{\Gamma}(z,h)|z_2, \ldots, z_T) = \Gamma(z) + d_K\Lambda(z)h^4 + o_p(h^4), \tag{2.8}$$

where $d_K = \int u^4 K(u)du - \sigma_K^4$. As a result, our criterion for bandwidth choice is

$$R(z,h) = u(z)^\top \hat{\Gamma}(z,h)u(z)(1 + g'(\theta(z))^2 V), \tag{2.9}$$

where V is the first diagonal element of $(Z'WZ)^{-1}(Z'W^2Z)(Z'WZ)^{-1}$, and $u(z) = (1 + \theta^2(z), -\theta(z))^\top$. Denote the minimizer of $R(z,h)$ as $\bar{h}$. Following

Fan and Gijbels (1995), one can show that $adj_K\bar{h}$ offers a reasonable approximation for $\hat{h}^{opt}$ in practice, where

$$adj_K = \left(\frac{4c_K d_K}{R(K)}\right)^{1/5} = 4^{1/5}\left(\frac{\int u^4 K(u)du}{\left(\int u^2 K(u)du\right)^2} - 1\right)^{1/5}.$$

To see this, using Fan and Gijbels (1995), we have

$$V = \frac{R(K)}{Thp(z)}(1 + o_p(1)). \tag{2.10}$$

It then follows from (2.9) and (2.10) that

$$\begin{aligned} E(R(z,h)|z_2,\ldots,z_T) = {} & u(z)^\top \Gamma(z)u(z) + d_K u(z)^\top \Lambda(z)u(z)h^4 \\ & + R(K)\frac{u(z)^T\Gamma(z)u(z)g'(\theta(z))^2}{Thp(z)} + o_p(h^4 + \frac{1}{Th}). \end{aligned}$$

It can be shown that the minimizer of the leading term of the above expression is $\hat{h}_o = \hat{h}^{opt}/adj_K$.

As $R(z,h)$ depends on the unknown $\theta(z)$, one can use $\hat{\theta}(z)$ with a prior bandwidth h to replace $\theta(z)$. The constant adj_K is determined by the chosen kernel function, for example, $adj_K = (92/7)^{1/5}$ for the Epanechnikov kernel.

Our practical choice of bandwidth is then

$$\begin{aligned} \tilde{h}^{opt}(z) &= adj_K \times \text{argmin}_{h\in\mathcal{H}}\hat{R}(z,h) \\ &= adj_K \times \text{argmin}_{h\in\mathcal{H}}\hat{u}(z)^\top\hat{\Gamma}(z,h)\hat{u}(z)(1 + g'(\hat{\theta}(z))^2 V), \end{aligned} \tag{2.11}$$

where $\hat{u}(z) = (1 + \hat{\theta}^2(z), -\hat{\theta}(z))^\top$. To obtain a globally optimal bandwidth, one can minimize $IR(h) = \int \hat{R}(z,h)dz$ and use $adj_K \cdot argmin_h IR(h)$ as the bandwidth. For implementation, the integral can be approximated by a discrete summation over the observed data. Undersmoothing is often desired as one would like to avoid biased estimation.

2.4. Model specification test

When the coefficient function $\theta(z)$ is a constant, using an FMA model can result in a loss in estimation efficiency; when it is not, using a misspecified MA(1) model can produce erroneous inference. A model specification test is needed to check if the specification of the FMA model is adequate.

We adopt the L2 norm-based test for regression functions (degenerated to a parameter in our case) proposed by Härdle and Mammen (1993) for testing the constancy of $\theta(z)$, due to its simple implementation. Thus one test H_0 against H_1, where

$$H_0 : P(\theta(z) \equiv \theta \text{ for some } \theta \in \mathbb{R}) = 1,$$

$$H_1 : P(\theta(z) \equiv \theta \text{ for some } \theta \in \mathbb{R}) < 1.$$

Similar to Härdle and Mammen's approach, we consider

$$D_T = Th^{1/2} \int_R (\hat{\theta}(z) - \hat{\theta})^2 \pi(z) dz,$$

where $\hat{\theta}$ is either the maximum likelihood estimator (MLE) if we assume Gaussian innovations, or the pseudo MLE if we do not assume Gaussian innovations, under H_0. Our test statistic does not have a smoothing operator on the parametric part, as advocated in Härdle and Mammen's (1993) test.

Let $m_{0,0}(z) = E\{(x_t - \mu)^4 | z_t = z\}$, $m_{1,0}(z) = m_{0,1}(z) = E\{(x_t - \mu)^3 (x_{t-1} - \mu)\} | z_t = z)$ and $m_{11}(z) = E\{(x_t - \mu)^2 (x_{t-1} - \mu)^2\} | z_t = z)$. We need the following assumption in additional to (A1)-(A6) for the specification test considered in this section.

(A7) The conditional moment functions $m_{i,j}(z)$, for $i, j = 0$ and 1, have continuous first derivatives at z for any $z \in R$.

Let $\hat{a}(z) = (\hat{a}_1(z), \hat{a}_0(z))$, $a(z) = (a_1(z), a_0(z))$, and write $\hat{\theta}(z) = h[g\{\hat{a}_1(z)/\hat{a}_0(z)\}] =: q\{\hat{a}_1(z), \hat{a}_0(z)\}$, and $\partial q(z)/\vec{a} = (\partial q(z)/a_1, \partial q(z)/a_0)^T$. Let

$$\mu_T = h^{-1/2} \int \{g^{'}\{\theta(z)\}^2 \nu(z)\big] \pi(z) dz,$$

$$\sigma_T^2 = 2K^{(4)}(0) \int f^{-2}(z) \big[\sum_{i,j=0}^{1} \frac{\partial q(z)}{\partial a_i} \frac{\partial q(z)}{\partial a_j} m_{i,j}(z) \big]^2 \pi^2(z) dz.$$

Theorem 4. *Under* (A1)∼(A7)*, then under* H_0

$$\sigma_T^{-1}(D_T - \mu_T) \xrightarrow{d} N(0, 1);$$

under H_1*,* $D_T \xrightarrow{p} \infty$ *at the rate of* $Th^1/2$*.*

Let $\hat{\mu}_{T0}$ and $\hat{\sigma}_{T0}$ be consistent estimators of μ_T and σ_T under H_0; these can be obtained by substituting estimates for the unknown quantities. As we advocate for a bootstrap implementation of the test, the detailed forms of $\hat{\mu}_{T0}$ and $\hat{\sigma}_{T0}$ are not important.

A test with nominal α level of significance for H_0 versus H_1 rejects H_0 if

$$D_T \geq \hat{\mu}_T + \hat{\sigma}_T z_{1-\alpha},$$

where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of $N(0, 1)$. Theorem 4 shows that the test has asymptotic size α, while power converges to 1 as $T \to \infty$. Hence the test is consistent.

To approximate the finite sample distribution of D_T under H_0, we use parametric bootstrap method in the spirit of Chen and Gao (2007):

Step 1. Apply the MA(1) model to x_t and obtain the estimator of the mean, $\hat{\mu}$, the coefficient, $\hat{\theta}$ and the variance, $\hat{\sigma}^2$.

Step 2. Generate a bootstrap re-sample according to $x_t^* = \hat{\mu} + \epsilon_t^* + \hat{\theta}\epsilon_{t-1}^*$ for $t = 1, 2, \ldots, T$, where $\{\epsilon_t^*\}_{1 \le t \le T}$ are independent $N(0, \hat{\sigma}^2)$ variables and obtain an estimate $\hat{\theta}(z)$ based on the resample.

Step 3. Repeat Step 2 B times for a large integer B and obtain $\{\hat{\theta}^{(i)}(z)\}_{i=1}^B$.

Step 4. Calculate

$$D_T^{(i)} = Th^{1/2} \int_R (\hat{\theta}^{(i)}(z) - \hat{\theta})^2 \pi(z) dz, \ i = 1, 2, \ldots, B,$$

and take the $(1-\alpha)$th quantile of $\{D_T^{(i)}\}_{1 \le i \le B}$ as the critical value for the test.

For simplicity, one can set $\pi(z) = 1$ and use the discrete sum to approximate D_T. In the next section, we use numerical simulations to study the size and the power of the proposed test.

Although we have demonstrated the use of the Härdle-Mammen test formulation, alternative test procedures may be considered; for instance the ones based on Fan and Li (1996), or the empirical likelihood as advocated in Chen and Gao (2007, 2011) and Chen and van Keilegom (2009).

3. Finite Sample Investigation

We generated the state variable z_t from the ARIMA(1,0,1) process:

$$(1 - 0.5B)z_t = (1 + 0.5B)u_t,$$

where $\{u_t\}$ is *i.i.d.* standard normal. The response x_t was generated according to

$$x_t = \epsilon_t + s\theta(z_t)\epsilon_{t-1}$$

for some $s \in [0, 1]$, with $\{\epsilon_t\}$ *i.i.d.* standard normal and independent of $\{u_t\}$. Three functions chosen for $\theta(\cdot)$ were (1) $\theta_1(z) = 2e^{-z^2} - 1$; (2) $\theta_2(z) = sin(3z)$; (3) $\theta_3(z) = (e^{2z} - 1)/(e^{2z} + 1)$. These functions were selected to describe the common features of humped, oscillating and monotone functional forms.

3.1. Performance of estimation

Our estimator has a slower convergence rate when $\theta(z) = \pm 1$, so we only consider the case $\theta(z) < 1$ for the finite sample study. To do so, we shrank the chosen functions by setting $s = 0.8$. For each choice of $\theta(z)$ and each $T \in \{100, 200\}$, we generated $\{x_t, z_t\}_{t=1}^T$ for 500 samples and obtained estimates of $\theta(z)$ for each generated sample, the mean value of which is plotted in solid line in

Table 1. Average RMSE comparison.

	FMA		ARFMA	
	$T = 100$	$T = 200$	$T = 100$	$T = 200$
$\theta_1(z)$	0.22	0.12	0.37	0.15
$\theta_2(z)$	0.34	0.20	0.51	0.27
$\theta_3(z)$	0.31	0.12	0.36	0.16

Figures 2 to Figure 4, corresponding to $\theta_1(z)$, $\theta_2(z)$, and $\theta_3(z)$, respectively. The dashed line in each figure represents the true function $0.8 \cdot \theta_i(z)(i = 1, 2, 3)$ and the dotted lines are the (point-wise) mean value plus and minus the standard deviation. The top two figures plot our estimation results and the bottom two figures plot those of Wang (2008). The bandwidth choice for Wang's estimator follows his GCV criterion. Both methods perform well in small samples. To compare them quantitatively, we calculated the average root of mean square error (RMSE), i.e.

$$ARMSE = \frac{1}{n}\sum_{i=1}^{n} RMSE(\hat{\theta}(z_{(i)})),$$

where $z_{(1)}, \ldots, z_{(n)}$ is a pre-specified sequence equally spaced in $[-2.5, 2.5]$ with step size 0.25. It is seen from Table 1 that our estimator has lower ARMSE than that of Wang (2008). We note, in addition, that Wang's method is computational demanding due to its iterative nature. To be precise, if we estimate $\theta(z)$ at n points, $z_{(1)}, \ldots, z_{(n)}$, and the sample size is T, then Wang's method requires $O(mT + n)$ weighted regressions where m is the number of iteration steps (in each step, it requires to re-estimation of $\theta(z_t)$ for each $t = 1, \ldots, T$, and, in the last step, it requires estimation of $\theta(z_{(i)})$ for each $i = 1, \ldots, n$), while our method only requires $O(n)$ weighted regressions.

3.2. Finite sample distribution

We approximated the distribution of $\hat{\theta}(z)$ by simulations. Theorem 3 indicates that the asymptotic distribution of $\hat{\theta}(z)$ is determined by whether the true value lies on the boundary or not. We treat these two cases separately. For each $T = 100, 200$, we generated a sample $\{x_t, z_t\}_{t \le T}$ for 500 times, and obtained 500 estimates of $\theta(z)$, denoted by $\hat{\theta}^{(1)}(z), \hat{\theta}^{(2)}(z), \ldots, \hat{\theta}^{(500)}(z)$. Their kernel density was calculated and compared to the asymptotic distribution of $\hat{\theta}(z)$.

When $\theta(z) = \pm 1$, the asymptotic distribution function of $\hat{\theta}(z)$ is discrete at ± 1, with the size of the atom being $1/2$ at the origin. Even if $|\theta(z)| < 1$, there are still some estimates concentrating on ± 1 when the sample size is not large enough. Thus, if we use kernel density as the empirical density, there might be two peaks at -1 and 1, which is not desirable for comparison. To circumvent this, we turned

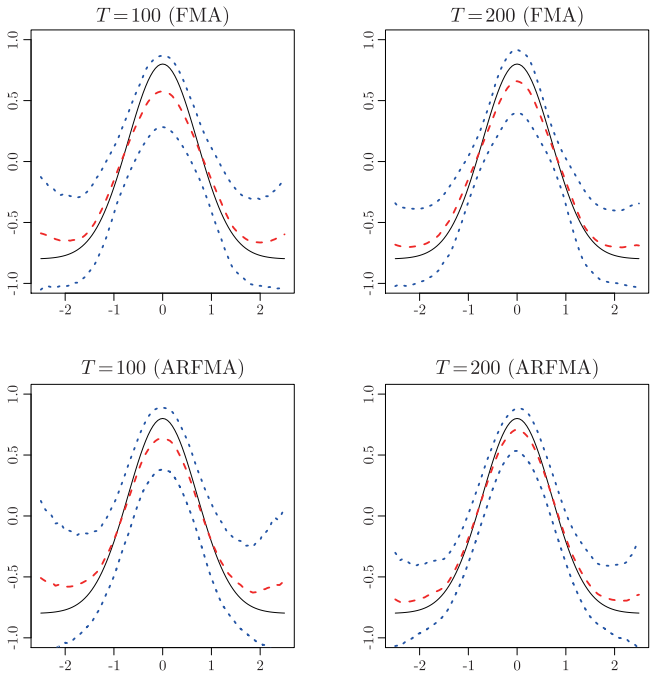

Figure 2. Plot of the true function $0.8 \cdot \theta_1(z)$ (dashed lines), averaged estimates (solid lines) and the associated one standard deviation confidence bands (dotted lines).

to the asymptotic conditional distribution of $\sqrt{Th/\nu(z)}(\hat{\theta}(z) - \theta(z) - G(z)h^2)$ given $|\hat{\theta}(z)| < 1$. When $|\theta(z)| < 1$, this distribution function is $\Phi(z)$ since $P(|\hat{\theta}(z)| < 1) \to 1$; when $\theta(z) = 1$, the conditional distribution is $2\Phi(-z^2/4)$ for $z \in (-\infty, 0)$; when $\theta(z) = -1$, the conditional distribution is $2\Phi(-z^2/4)$ for $z \in (0, \infty)$. We compared the kernel density of $\{\hat{\theta}^{(i)}(z) : |\hat{\theta}^{(i)}(z)| < 1\}$, to the corresponding asymptotic distribution. In addition, we also computed the fraction of times that $|\hat{\theta}^{(i)}(z)| = 1$, denoted by $P(A)$. This should be close to 0 when $|\theta(z)| < 1$ and 0.5 when $\theta(z) = \pm 1$ for large enough T.

We set $s = 1$ and $\theta(z) = \theta_1(z)$ to illustrate the findings. First, consider the estimation of $\theta(z)$ at $z_0 = \sqrt{\log 2}$. Here $\theta(z_0) = 0 \in (-1, 1)$. The empirical conditional densities of the standardized data are plotted in Figure 5 and the probability $P(A)$ is reported at the bottom of each subfigure. The bandwidth of kernel density was selected by cross validation. The dashed line is the standard normal density and the solid line is the kernel density. The lines are close to each other even for moderate T and $P(A)$ decreases to 0 as the sample size grows.

To study the boundary issue, we estimate $\theta(z)$ at $z_0 = 0$ ($\theta(z_0) = 1$). The conditional kernel densities of the standardized data are plotted in Figure 6. The

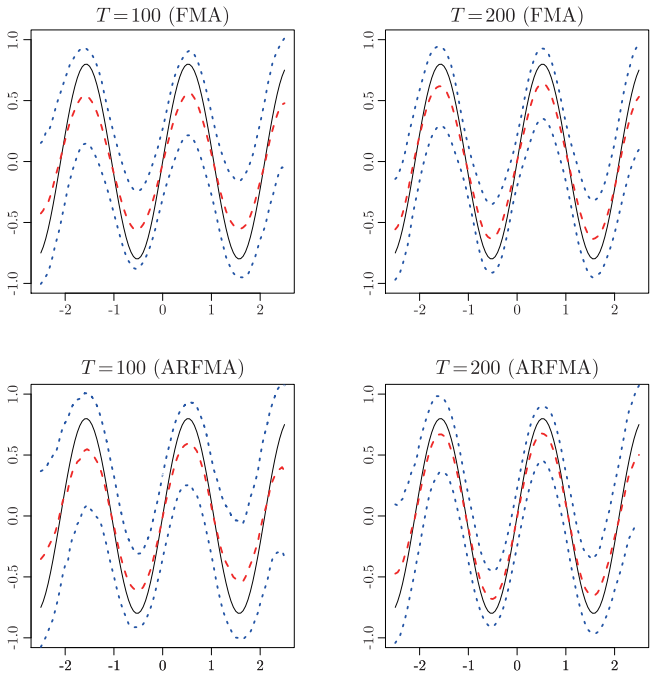

Figure 3. Plot of the true function $0.8 \cdot \theta_2(z)$ (dashed lines), averaged estimates (solid lines) and the associated one standard deviation confidence bands (dotted lines).

Table 2. Bias and variance of $\hat{\theta}(z)$ at z_0: $\theta(z) = 2e^{-z^2} - 1$.

	T	100	200	500	1,000
$z_0 = \sqrt{\log 2}$	Bias	-0.29	-0.24	-0.18	-0.17
	Var	0.30	0.25	0.21	0.19
$z_0 = 0$	Bias	0.01	0.02	0.01	0.01
	Var	0.35	0.23	0.15	0.10

lines are close to each other even for moderate T, and $P(A)$ increases to 0.5 as the sample size increases.

The bias and variance of the estimator $\hat{\theta}(z)$, for $z_0 = \sqrt{\log 2}$ and $z_0 = 0$, are reported in Table 2. It can be seen that the variance decreases as the sample size T increases. The bias is decreasing in the first case, while remaining close to 0 in the second case. This is because that the asymptotic bias, $G(z)h^2$, is diminishing in the first case, while it is always 0 in the second case.

3.3. Size and power of the test

In this subsection, we consider the size and the power of the model specifi-

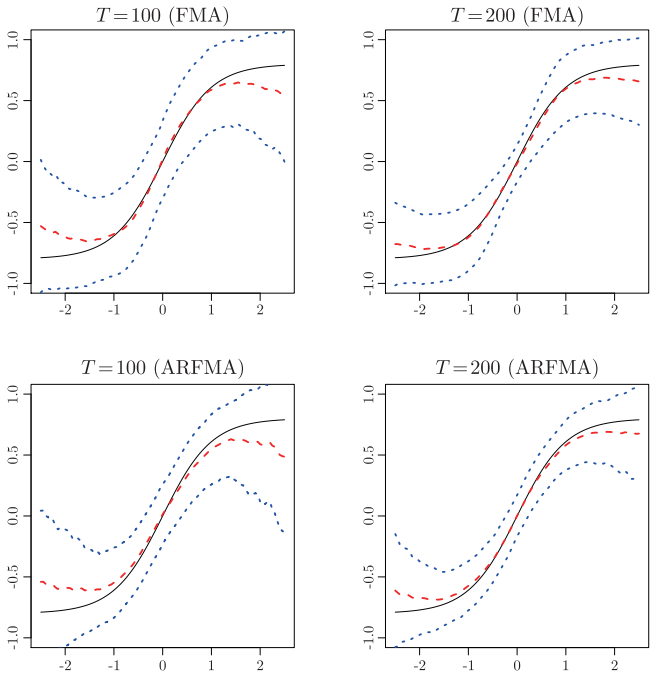

Figure 4. Plot of the true function $0.8 \cdot \theta_3(z)$ (dashed lines), averaged estimates (solid lines) and the associated one standard deviation confidence bands (dotted lines).

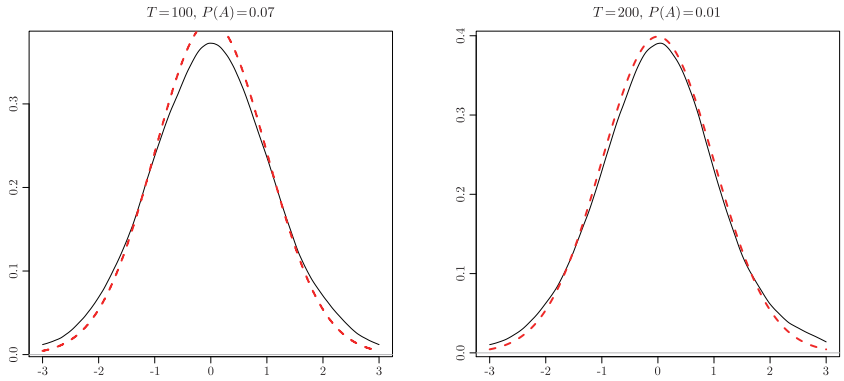

Figure 5. The finite sample distribution of $\hat{\theta}_1(z)$ at $z_0 = \sqrt{\log 2}$ (solid lines) and the theoretical asymptotic distribution (dashed lines) together with the probability of $A = \{\hat{\theta}_1(z_0) = \pm 1\}$.

cation test via simulation. The size was estimated by the proportion of rejection under the null hypothesis while the power was estimated by that under the alternative. As for the size, we consider the DGP $x_t = \epsilon_t + \theta\epsilon_{t-1}$, where ϵ_t's were

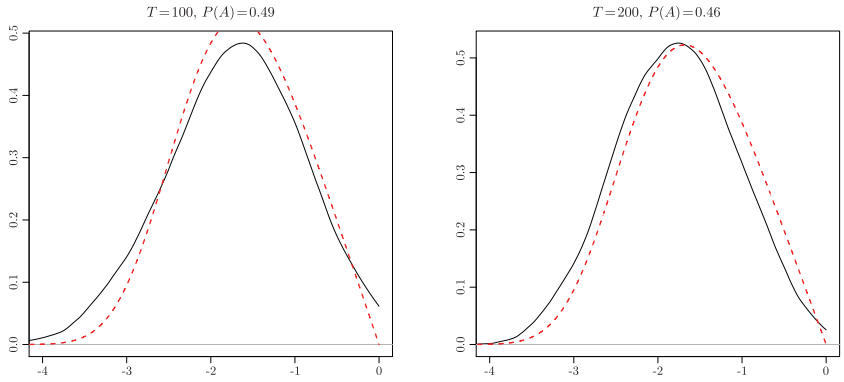

Figure 6. The finite sample conditional distribution of $\hat{\theta}_1(z)$ at $z_0 = 0$ given $|\hat{\theta}_1(z_0)| < 1$ (solid lines) and the theoretical asymptotic distribution (dashed lines) together with the probability of $A = \{\hat{\theta}_1(z_0) = \pm 1\}$.

Table 3. Rejection rate (%) Under H_0 ($\alpha = 5\%$).

T	$\theta = 0.2$	0.4	0.6	0.8	1.0
100	3.0	4.4	3.4	4.2	5.0
200	5.0	4.2	4.2	5.0	4.6

i.i.d. standard normal, with θ set to be 0.2, 0.4, 0.6, 0.8, and 1.0, respectively. For each θ and each sample size $T \in \{100, 200\}$, we generated 500 sets of data and calculated the proportion of rejection when the significance level α is 5%, with bootstrap resamples $B = 100$. The bandwidth was selected based on the RSC criterion. The results are reported in Table 3. It can be seen that our test has proper size.

As for the power, we considered the DGPs $x_t = \epsilon_t + s \cdot \theta_j(z_t)\epsilon_{t-1}$, where ϵ_t's were *i.i.d.* standard normal, $j \in \{1, 2, 3\}$ and $s \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$. For each design and sample size $T \in \{100, 200\}$, we generated 500 sets of data and calculated the proportion of rejection when the significance level α is 5%, with bootstrap resamples $B = 100$. The bandwidth was selected based on the RSC criterion. The results are reported in Table 4 and it is seen that the rejection rate gets larger as s increases. For moderate value of s, the power is decent.

4. Data Analysis

4.1. Application to the Chinese CPI

We applied an FMA model to Chinese CPI data and compared its forecast performance to that of MA model. The year-on-year CPI monthly growth data ranging from January 1990 to March 2014 was downloaded from Wind database (`www.wind.com.cn`). The raw data is plotted in panel (a) of Figure 7. It is clear

Table 4. Rejection rate (%) Under H_1 ($\alpha = 5\%$).

$\theta(z)$	T	$s = 0.2$	0.4	0.6	0.8	1.0
$\theta_1(z)$	100	6.0	20.0	36.2	54.2	68.0
	200	13.4	56.2	85.4	96.0	97.6
$\theta_2(z)$	100	6.2	14.6	30.8	57.0	64.2
	200	10.0	34.4	71.6	89.4	95.6
$\theta_3(z)$	100	11.2	38.4	70.8	79.0	91.2
	200	27.4	75.6	97.8	100.0	100.0

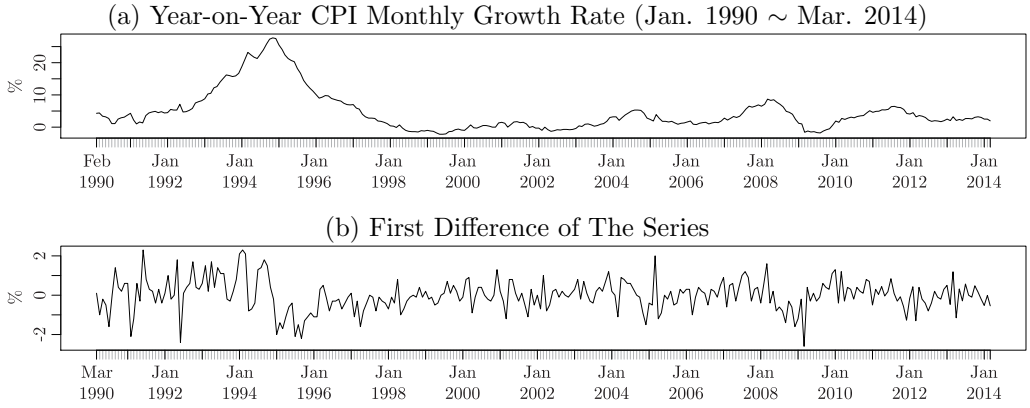

Figure 7. (a) The Chinese CPI monthly growth rate and (b) its first order difference, Jan. 1990 to Mar. 2014.

that the data is nonstationary, as also supported by the ADF test. The first order difference of the data is plotted Figure 7(b), which is found stationary.

Our target is to forecast the data ranging from January 2011 to March 2014. When the MA(1) model was used for the first-order differenced log of CPI (or equivalently, ARIMA(0,1,1) for logCPI), the root mean squared forecast error (RMSE) was computed as 0.589.

For forecasts using FMA(1), the first set of state variables considered were various measures of money supply, including M0, M1, and M2, as the neutrality of money implies that an increase in money supply will eventually convert to an increase in price level. Other economic variables that can affect the level of price, including export (Ex), import (Im), retail sales (RS) and PPI were also considered. Since PPI is often presumed to be the leading index of CPI, we also considered 4 sub-categories of PPI: capital goods (ca), consumer goods (co), light manufacturing (lm) and heavy manufacturing (hm). Year-on-year growth rate data for these 11 state variables were obtained from Wind Database. Since all variables are I(1), first-order differenced series were used.

First, we conducted the model specification test to detect the state variables whose corresponding coefficient functions differ significantly from a constant.

Table 5. Significant state variables.

z_{t-d}	d	z_{t-d}	d
M0	12	PPI	4, 11
M1	9	ca	11
M2	8, 9	co	12
Ex	2, 11, 12	lm	7, 8, 9, 12
Im	2, 12	hm	11
RS	11, 12		

Table 6. Forecasting RMSE of FMA(1) with various variables.

z_t	$\mathrm{M0}_{t-12}$	$\mathrm{M1}_{t-9}$	$\mathrm{M2}_{t-8}$	$\mathrm{M2}_{t-9}$
	0.524	0.572	0.638	0.525
z_t	Ex_{t-2}	Ex_{t-11}	Ex_{t-12}	Im_{t-2}
	0.532	0.505	0.563	0.549
z_t	Im_{t-12}	RS_{t-11}	RS_{t-12}	PPI_{t-4}
	0.463	0.521	0.575	0.607
z_t	PPI_{t-11}	ca_{t-11}	co_{t-12}	lm_{t-7}
	0.519	0.528	0.494	0.577
z_t	lm_{t-8}	lm_{t-9}	lm_{t-12}	hm_{t-11}
	0.513	0.502	0.508	0.543

For each variable, we included lagged variables starting from the 2nd order to the 12th order. The 1st order lagged variables were excluded for identification requirements. Among all of 121 state variables (11 variables with 11 lags for each), we found 20 of them significant at level $\alpha = 5\%$. These findings are summarized in Table 5.

We plot the estimates of $\theta(\cdot)$ for the choice of 6 significant variables, $M0_{t-12}$, $M2_{t-8}$, Ex_{t-11}, Im_{t-12}, co_{t-12}, and lm_{t-12}, as illustrated in Figure 8; they display strong departure from constancy.

The forecast RMSEs using different state covariates are summarized in Table 6. Among these 20 state variables, over 85% of them outperformed the MA(1) model in terms of the forecast RMSE. Specifically, we found that the 12th lag of import led to the best forecasts. The forecast RMSE was 0.463, which is a 21.4% improvement over that of the MA(1) model.

4.2. Application to german egg price

We also analyzed the German egg data (Finkenstadt (1995); Fan and Yao (2003); Wang (2008)) using our FMA model. To compare it with the ARFMA model of Wang (2008), we similarly fit the first 290 data points, leaving out the next 10 points for assessing the prediction accuracy. Figure 9 displays the series as well as its autocorrelation function (ACF) and partial autocorrelation function

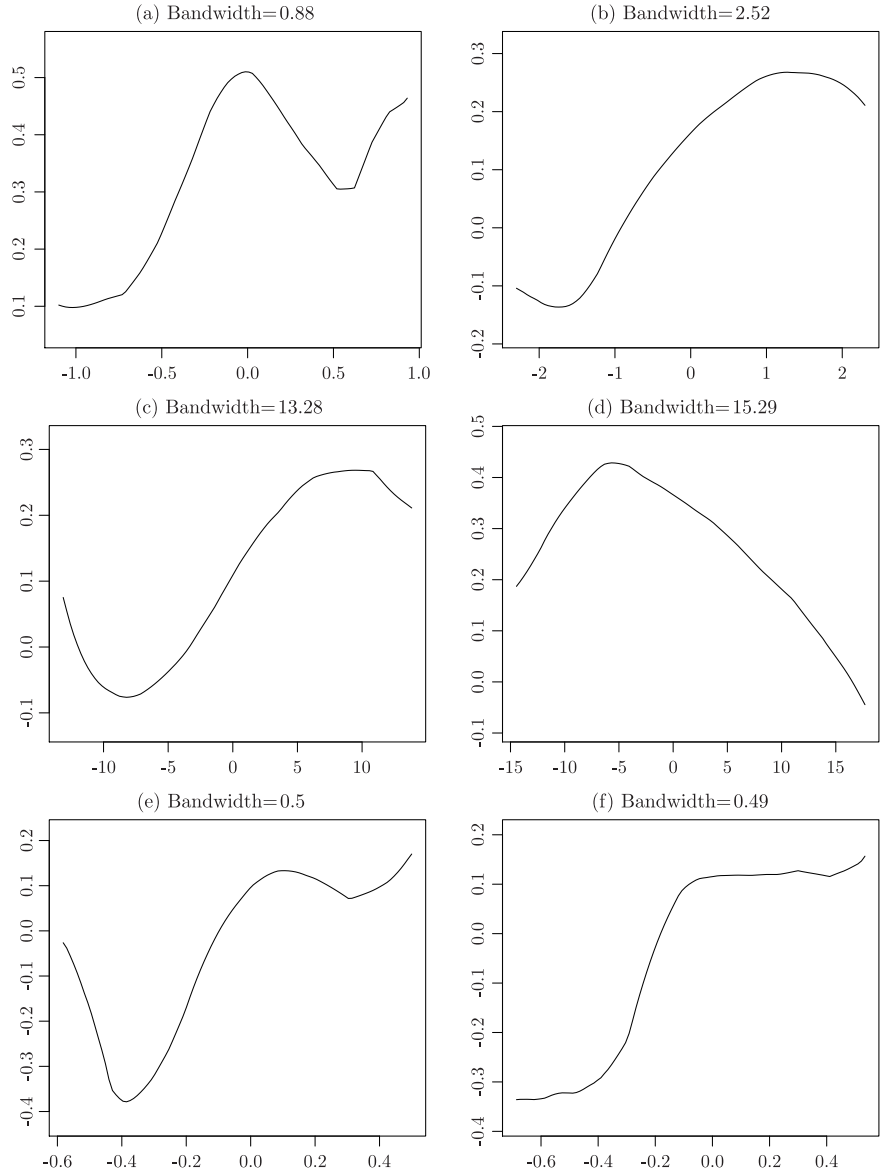

Figure 8. Estimates of $\theta(z_t)$ where (a) z_t=$\Delta \mathrm{M0}_{t-12}$; (b) z_t=$\Delta \mathrm{M2}_{t-8}$; (c) z_t=Ex_{t-11}; (d) z_t=Im_{t-12}; (e) z_t=co_{t-12}; (f) z_t=lm_{t-12}.

(PACF). It can be observed that the ACF varies slowly while PACF decreases quickly. After fitting an AR model to the first 290 data points, we found that the original series is stationary but the AR(1) coefficient was 0.95. Thus, we transformed the original series P_t by $X_t = P_t - 0.95P_{t-1}$ to eliminate the strong dependency. Figure 10 displays the transformed series, its ACF and PACF. The

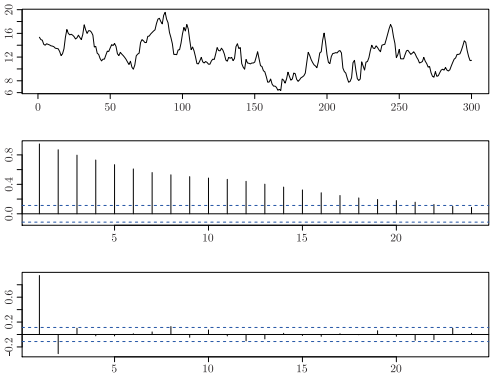

Figure 9. Egg Price: Series (upper), ACF (middle) and PACF (lower).

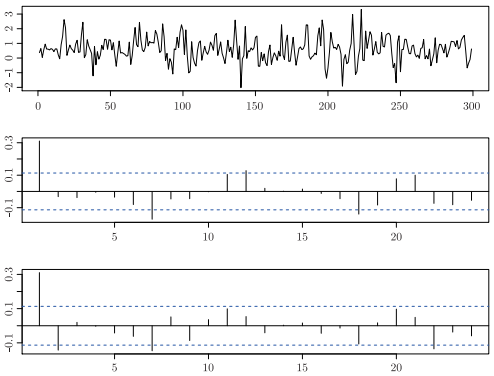

Figure 10. Transformed Egg Price: Series (upper), ACF (middle) and PACF (lower).

ACF plot suggests that an FMA(7) may be used.

First, we fit an FMA(1) model on X_t with state covariate X_{t-d},

$$X_t = \epsilon_t + \theta(X_{t-d})\epsilon_{t-1}.$$

Similar to Wang (2008), only $1 \le d \le 5$ was considered. The bandwidth was selected based on the procedure mentioned in Section 2.3. We found that $d = 3$ resulted in the best prediction performance as measured in average absolute prediction error (AAPE), following Wang (2008). The AAPE for the FMA(1) model with $d = 3$ was 0.477, 7.7% better than Wang's (2008) best prediction. The estimation of $\theta(X_{t-3})$ is displayed in Figure 11, apparently non-constant.

We consider higher order FMA models as well. As with Wang (2008), we considered the class of models

$$X_t = \epsilon_t + \theta_1(X_{t-d})\epsilon_{t-1} + \theta_k(X_{t-d})\epsilon_{t-k}.$$

When $k \ge 3$, the conditional autocorrelation can be written as

$$\begin{aligned} E(X_t^2|X_{t-d} = z) &= (1 + \theta_1(z)^2 + \theta_k(z)^2)\sigma^2; \\ E(X_tX_{t-1}|X_{t-d} = z) &= \theta_1(z)\sigma^2; \\ E(X_tX_{t-k}|X_{t-d} = z) &= \theta_k(z)\sigma^2. \end{aligned}$$

Therefore, one can use similar methods to estimate $\theta_1(z)$ and $\theta_k(z)$ as in the FMA(1) model. The bandwidth was also selected based on RSC criterion described in Section 2.3. Among all models, we found that $d = 4$, $k = 7$ led to the best prediction. The corresponding AAPE was computed as 0.439, which further improves that of FMA(1) by 7.5%. The estimates of $\theta_1(X_{t-4})$ and $\theta_7(X_{t-4})$ are plotted in Figure 12. We conclude that our proposed method does better than that of Wang (2008).

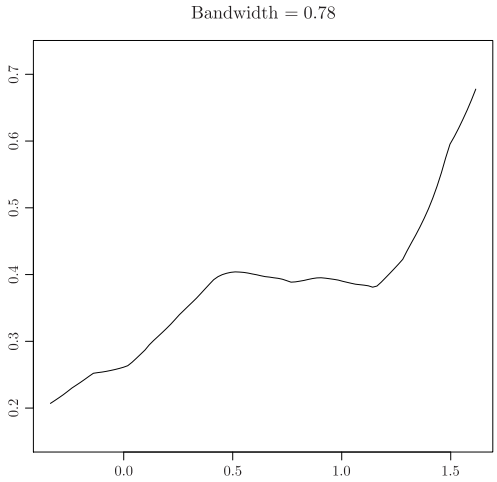

Figure 11. Estimates of $\theta_1(z)$ in FMA(1).

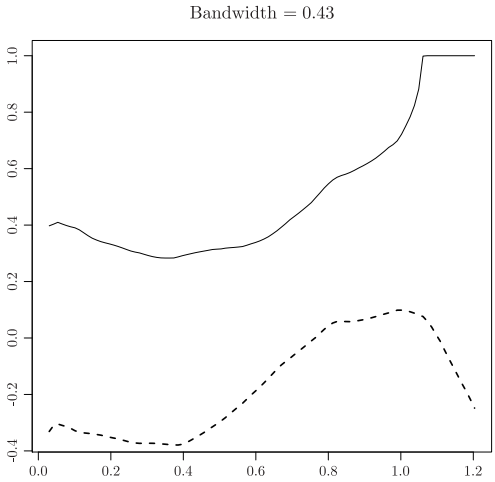

Figure 12. Estimates of $\theta_1(z)$ (solid line) and $\theta_7(z)$ (dashed line) in FMA(7).

5. Conclusion

This paper extends moving averaging models by allowing the MA coefficients to adapt with a covariate. Under parameter identification, we proposed to estimate the functional coefficient by a ratio of two conditional moment estimators derived from local linear least squares. The consistency and asymptotic distribution of the proposed estimators are established. A Härdle and Mammen type adequacy test of the constancy of the functional coefficient is also proposed. Both simulation and empirical exercises show that our proposed method perform well in finite samples.

The FMA(1) framework can be extended to the general ARFMA(p,q). We outline how the extension can be made via the ARFMA(1,2)

$$x_t - \alpha x_{t-1} = \epsilon_t + \theta_1(z_t, z_{t-1})\epsilon_{t-1} + \theta_2(z_t, z_{t-1})\epsilon_{t-2}, \tag{5.1}$$

where α is the AR coefficient, and $\theta_1(\cdot)$ and $\theta_2(\cdot)$ are two MA nonparametric coefficient functions that depend on (z_t, z_{t-1}), as suggested by a referee. We take the mean of x_t to be zero in (5.1) to simplify the notation. After algebraic manipulation, similar to those exhibated in (3)−(4), it can be shown that

$$\begin{aligned} &Var(x_t|z_t, z_{t-1}) - 2\alpha\, Cov(x_t, x_{t-1}|z_t, z_{t-1}) + \alpha^2 Var(x_{t-1}|z_t, z_{t-1}) \\ &\quad = \sigma^2\{1 + \theta_1^2(z_t, z_{t-1}) + \theta_2^2(z_t, z_{t-1})\}, \qquad (5.2) \\ &Cov(x_t, x_{t-1}|z_t, z_{t-1}, z_{t-2}) - \alpha Var(x_{t-1}|z_t, z_{t-1}, z_{t-2}) \\ &\quad = \sigma^2\{\theta_1^(z_t, z_{t-1}) + \theta_1(z_{t-1}, z_{t-2})\theta_2(z_t, z_{t-1})\}, \qquad (5.3) \end{aligned}$$

$$Cov(x_t, x_{t-2}|z_t, z_{t-1}) - \alpha\, Cov(x_{t-1}, x_{t-2}|z_t, z_{t-1}) = \sigma^2\theta_2(z_t, z_{t-1}), \quad (5.4)$$

$$Cov(x_t, x_{t-3}|z_t, z_{t-1}) - \alpha\, Cov(x_{t-1}, x_{t-3}|z_t, z_{t-1}) = 0. \quad (5.5)$$

Let $g_j(z_1, z_2) = Cov(x_t, x_{t-j}|z_t = z_1, z_{t-1} = z_2)$ for $j = 0, 1, 2, 3$, $g_{3+j}(z_1, z_2) = Cov(x_{t-1}, x_{t-j}|z_t = z_1, z_{t-1} = z_2)$ for $j = 1, 2, 3$, and $g_{7+j}(z_1, z_2, z_3) = Cov(x_{t-j}, x_{t-1}|z_t = z_1, z_{t-1} = z_2, z_{t-2} = z_3)$. Carry out the local linear estimation for these functions, and denote the estimator as $\hat{g}_k(z_1, z_2)$ for $k = 0, 1, \cdots$ and 8. Then, an estimator for α is

$$\hat{\alpha} = \frac{n^{-1}\sum_{t=1}^{n} \hat{g}_3(z_t, z_{t-1})}{\hat{g}_6(z_t, z_{t-1})},$$

which should be more efficient than having the estimation based on a single or a few (z_t, z_{t-1}). The estimators for $\theta_1(z_1, z_2)$ and $\theta_2(z_1, z_2)$ can be obtained by solving the estimating equations based on (5.2) to (5.5). The conditions assumed for FMA(1) given in Assumptions (A2)-(A5) in Section 2.2 need to be updated by replacing z_t by the pair (z_t, z_{t-1}, z_{t-2}).

We can see that as the order of the ARFMA increases, the estimation procedure involves more functions. Hence, ARFMA(p, q) models with lower order are more practically useful. Indeed, a criterion one should adopt in choosing the state covariate z_t is that it allows shorter orders in the ARFMA(p, q). Stationarity conditions for ARFMA would follow as in the discussion of Wang (2008). Alternatively, a semiparametric single index structure may be imposed for the smooth coefficient function to allow for multiple state variables. There is more to investigate on these topics.

Supplementary Materials

The online supplementary materials contain some useful lemmas and the proofs of the main theorems.

Acknowledgement

The authors thank Editor Qiwei Yao, an associate editor, two anonymous referees, and seminar participants at School of Statistics and Management, Shanghai University of Finance and Economics, China Meeting of Econometric Society 2014, and Financial Engineering and Risk Management 2014 for helpful comments and suggestions. We acknowledge support from National Statistics Bureau, Center for Statistical Science and LMEQF at Peking University. Tu (the corresponding author) acknowledges support from National Natural Science Foundation of China Grants 71301004, 71472007 and 71532001. This work grows out of the weekly discussion of Yandong-School of Data and an earlier draft of this paper was Lei's undergraduate thesis under Chen and Tu's supervision.

References

Barsky, R. B. (1987). The Fisher Hypothesis and the forecastability and persistence of inflation. *J. Monetary Econom.* **19**, 3-24.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *J. Econom.* **31**, 307-327.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Revised edition, Holden Day, San Francisco.

Box, G. E. P. and Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *J. Amer. Statist. Assoc.* **70**, 70-79.

Cai, Z. and Li, Q. (2008). Nonparametric estimation of varying coefficient dynamic panel data models. *Econom. Theory* **24**, 1321-1342.

Cai, Z., Fan, J. and Yao, Q. (2000). Functional-Coefficient Regression Models for Nonlinear Time Series. *J. Amer. Statist. Assoc.* **95**, 941-956.

Caner, M. and Hansen, B. (2001). Threshold autoregression with a unit root. *Econometrica* **69**, 1555-1596.

Chan, K. S. and Tong H. (1986). On estimating thresholds in autoregressive models. *J. Time Series Anal.* **7**, 179-190.

Chen, R. and Tsay, R. S. (1993). Functional-coefficient autoregressive models. *J. Amer. Statist. Assoc.* **88**, 298-308.

Chen, S. X. and Gao, J. (2007). An adaptive empirical likelihood test for time series models. *J. Econom.* **141**, 950-972.

Chen, S. X. and Gao, J. (2011). Simultaneous specification test for the mean and variance structures for nonlinear time series regression. *Econom. Theory* **27**, 792-843.

Chen, S. X., Gao, J. and Tang, C. Y. (2008). A test for model specification of diffusion processes. *Ann. Statist.* 36, 167-198.

Chen, S. X. and van Keilegom, I. (2009). Empirical likelihood test for a class of regression models. *Bernoulli*, **15**, 955-976.

Cleveland, W. P. and Tiao, G. C. (1976). Decomposition of seasonal time series - a model for the census X-11 program. *J. Amer. Statist. Assoc.* **71**, 581-587

Dahlhaus, R. (1989), Efficient parameter estimation for self-similar processes. *Ann. Statist.* **17**, 1749-1766.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica* **55**, 251-276.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc.* **57**, 371-394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series–Nonparametric and Parametric Methods*. Springer-Verlag, Berlin.

Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *J. Roy. Statist. Soc. Ser. B* **65**, 57-80.

Fan, Y. and Li, Q. (1996). Consistent model specification tests: omitted variables and semiparametric function forms. *Econometrica* **64**, 865-90.

Finkenstadt, B. (1995). *Nonliear Dynamics in Economics*. Springer, Berlin.

Granger, C. W. J. and Joyeux, R. (1980). An introductionto long- memory time series models and fractional differencing. *J. Time Series Anal.* **1**, 15-29.

Haggan, V. and Ozaki, T. (1981). Modeling nonlinear vibrations using an amplitude-dependent autoregressive time series model. *Biometrika* **68**, 189-196.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hannan, E. J. and Deistler, M. (1988), *The Statistical Theory of Linear Systems*. Wiley, New York.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21**, 1926-1947.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econom.* **58**, 71-120.

Masry, E. and Fan, J. (1997). Local polynomial estimation for stationary stable processes. *Stochastic Process. Appl.* **18**, 1-31.

Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9**, 141-142.

Nelson, C. R. and Schwert, G. W. (1977). Short-term interest rates as predictors of inflation: on testing the hypothesis that the real rate of interest is constant. *Amer. Econom. Rev.* **67**, 478-486.

Ozaki, T. (1982). The statistical analysis of perturbed limit cycle processes using nonlinear time series models. *J. Time Series Anal.* **3**, 29-41.

Priestley, M. B. (1980). State-dependent models: a general appraoch to nonlinear time series analysis . *J. Time Series Anal.* **1**, 47-71.

Stock, J. H. and Watson, M. W. (2006). Why has U.S. inflation become harder to forecast? NBER Working Paper no. 12324.

Schwert, G. W. (1987). Effects of model specification on tests for unit roots in macroeconomic data. *J. Monetary Econom.* **20**, 73-103.

Tiao, G. C. and Tsay, R. S. (1994). Some advances in nonlinear and adaptive modeling in time series. *J. Forecasting* **13**, 109-131.

Tong, H. (1983). Threshold autoregression, limit cycles and cyclical data(with discussion). *J. Roy. Statist. Soc. Ser. B* **42**, 245-292.

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

Wang, H. B. (2008). Nonlinear ARMA models with functional MA coefficients. *J. Time Series Anal.* **19**, 1032-1056.

Watson, G. S. (1964). Smooth regression analysis. *Indian J. Statist. A* **26**, 359-372.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1-26.

Guanghua School of Management and Center for Statistical Science, Peking University, Beijing, China, 100871.

Iowa State University, Ames, USA.

E-mail: csx@gsm.pku.edu.cn

School of Mathematics, Peking University, Beijing 100871, China.

E-mail: leilihuallh@126.com

Guanghua School of Management and Center for Statistical Science, Peking University, Beijing, 100871, China.

E-mail: yundong.tu@gsm.pku.edu.cn

(Received May 2014; accepted April 2015)