LINEAR REGRESSION WITH AR(∞) ERRORS UNDER CONSTRAINED COEFFICIENTS APPLYING MAXIMUM A POSTERIORI ESTIMATION

Ruzhang Zhao and Dong Li*

Tsinghua University

We discuss linear regression with autoregressive error and there are infinite coefficients for the AR error term. Also, coefficients of AR error term are restricted to an ellipsoid. At least, the coefficients should be absolutely summable. Here, we use estimator under constrainted norms. Consistency theorems for the coefficients of AR error are provided based on maximum a posteriori estimation. In particular, the consistency rate for the coefficients of AR error between the estimator and real value. In the previous work, sieve estimation of the coefficients requires the sample size goes to infinity in a controlled way, but under the restriction of ellipsoid, there is no need for such constraint. The estimation of AR error term under constrainted norms is shown to be robust by further empirical application.

1. Introduction. Regression with correlated error term is a widely used tool in applied economics. A leading topic is about what kind of correlation can be applied in the error term and provide good prediction. Cochrane and Orcutt (1949) used basic least squares regression to analysis regression with correlated error term. Sacks and Ylvisaker (1966) designed a similar way to solve the problem with correlated errors. Beach and MacKinnon (1978) used maximum likelihood method to consider the problem. And with the AR model put into the regression model, Chib (1993) used Bayesian method to analyze the AR error in the regression. Least squares estimation is still popular in Pierce (1971). The high order even infinite order of autoregressive models are considered in Bhlmann (1997). The sieve estimation for the infinite AR model is well explored and there are several consistent results for the sieve estimation.Bhlmann (1995, 1997) and Burman and Nolan (1992) discussed the consistency and application of the infinite AR model and get good results. Sancetta (2018) put forward a new way to solve the infinite AR model, and the coefficients are restricted into an infinite dimensional ellipsoid.

In the light of the arguments above, we consider high-order even infinite AR error term. According to constraint used in sieve estimation, there will be similar problems in the estimation of ARIMA or infinite AR model. The method used in Sancetta (2018) is illuminating to the analysis of infinite AR model.

Moreover, let us return to the researches about regression and maximum likelihood estimation. One can refer to Leggetter and Woodland (1995) and Lokshin et al. (2004) for comprehensive studies on regression with maximum likelihood estimation. Also, Bayes methods work well when being added into the maximum likelihood estimation. Furthermore, maximum a posteriori which is a kind of

*Supported in part by the NSFC (No.11571348, 11771239).

Keywords: AR Error, Consistency, Maximum A Posteriori Estimation, Linear Regression

maximum likelihood estimation is used as a powerful tool for problems about mixture model and DeGroot (2005) can be referred to for more details about maximum a posteriori estimation. Gauvain and Lee (1994), Nerurkar et al. (2009), Greig et al. (1989) and Doucet et al. (2002) applied maximum a posteriori and provided some useful results when estimating.

In the paper, we need to go further and allow for the case where coefficients of the AR error term and the regression coefficients are both required to estimate. When we consider the regression with the AR error or ARIMA error, which is hard to estimate the coefficients of the error without any constraint. So, constraints on the decay rate of the coefficients of the AR error are often considered for prediction. It is common to require the coefficients of AR error are absolutely summable, which is also the basic requirement for the consistent results of the coefficients. Schafer (2002) considers the consistency under such constraint.

In the paper, Bayesian Maximum A Posteriori Estimation will be applied for the estimation of the coefficients for infinite AR error term.

The paper is organized as follows. Section 2 outlines our basic coefficient estimation model. For better understanding the estimation, we expound our approach further about the ellipsoid and do some simple denotations for the future statement. Section 3 provides the estimator via maximum a posteriori estimation. Section 4 details how the estimation model works and provides a consistency result for the coefficients of our estimator. The following part of section proves the correctness of the theorems and lemmas.

2. Estimation Model. The basic linear regression model allows for some assumptions about the error term. The regression model is give by

$$Y_t = X_t^T \beta + \epsilon_t \tag{2.1}$$

where the error term $(\epsilon_t)_{t\in\mathbb{Z}}$ satisfies the condition of AR model whose coefficients are infinite.

$$\epsilon_t = \sum_{k=1}^{\infty} \phi_k \epsilon_{t-k} + u_t \tag{2.2}$$

where u_t's are independent identically distributed sequence with mean zero and variance $\sigma^2(>0)$ and satisfy the normal distribution, which means $u_t \sim N(0, \sigma^2)$. The coefficients ϕ_k's are unknown and the regression coefficient β is also unknown. In this paper, we estimate the coefficients under the condition that $\sum_{k=1}^{\infty} |\phi_k| \leqslant \bar{\phi} < \infty$, where the absolute summable condition is basic for consistency.

When we are given finite sample, we can use the finite dimensional model below to do approximation to the AR infinite error.

$$\epsilon_t = \sum_{k=1}^{K} \alpha_k \epsilon_{t-k} + u_t \tag{2.3}$$

where $K < \infty$ and α_k is a replace and approximation of ϕ_k. This is the normal process to do a sieve estimation. However, according to the process of sieve estimation, when we use the formula 2.3 to

approximate formula 2.2, we will have to make the sample size and the number of coefficients in 2.3 ,K controlled, and the approximation can be reasonable. However, here, we restrain the coefficients of the AR error in an ellipsoid with infinite dimensions. In the latter part of the paper, the advantages for using the ellipsoid can be shown, and because we restrain the coefficients, the sample size and the K above can go to infinite without any constraint. The ellipsoid is defined as follows

$$\Omega_K^p(A) := \{\alpha \in \mathbb{R}^\infty : \sum_{k=1}^{\infty} \alpha_k^p \lambda_k^p \leqslant A^p, \alpha_k = 0 \, for \, k > K\} \tag{2.4}$$

where $\lambda_k \propto k^\lambda$ for some $\lambda > 0$, $p \geq 2$ and $k \in \mathbb{N}$. Here, we use $\propto$ to represent that λ_k and k^λ are proportional and p is applied for p-norm.

Based on the definition of $\Omega_K(A)$, we can have the following three extensions.

$$\Omega^p(A) = \bigcup_{K>0} \Omega_K^p(A) \quad \Omega_K^p = \bigcup_{A<\infty} \Omega_K^p(A) \quad \Omega^p = \bigcup_{A<\infty} \Omega^p(A)$$

Moreover, we provide the following denotation.

REMARK 2.1. We denote the sample size as $N = n + K$ and the sample can be denoted as

$$Y_{-K+1}, Y_{-K+2}, \cdots, Y_0, \cdots, Y_n; X_{-K+1}, X_{-K+2}, \cdots, X_0, \cdots, X_n$$

And the sample size $N = n + K$ and the number of coefficients in 2.3 ,K can go to infinite without constraint except that $N > K$,which is clear from our denotation.

3. Estimator from maximizing a posteriori estimation. We first outline the estimator derived from least square estimation and the shortcoming for the method will be stated. Then the Bayesian Method- Maximum a Posteriori Estimation will be applied for linear regression model with infinite AR error term for better getting the consistency results. Based on the Maximum a Posteriori Estimation, we can consider the part of the coefficients instead of the whole coefficients. As for least square estimation, all the coefficients should be estimated.

The constrained problem can be considered when the coefficients lie in the ellipsoid $\Omega_K^p(A)$, and the constrained estimator is defined as

$$(\hat{\alpha}^n, \hat{\beta}) = \arg \inf_{\alpha \in \Omega_K^p(A), \beta} \frac{1}{n} \sum_{t=1}^{n} [(Y_t - X_t\beta) - \sum_{k=1}^{\infty} \alpha_k (Y_{t-k} - X_{t-k}\beta)]^2 \tag{3.1}$$

where we restrain the coefficient term α in the ellipsoid $\Omega_K^p(A)$, where only the first K elements of the vector α are non-zero. So, the estimator is the same as

$$(\hat{\alpha}^n, \hat{\beta}) = \arg \inf_{\alpha \in \Omega_K^p(A), \beta} \frac{1}{n} \sum_{t=1}^{n} [(Y_t - X_t\beta) - \sum_{k=1}^{K} \alpha_k (Y_{t-k} - X_{t-k}\beta)]^2 \tag{3.2}$$

According to the property of the least square estimation, if we have the estimator of β, the coefficient term α is easy to estimate. On the contrary, if we assume the coefficient term α is already known, β can be derived from 3.2. Assume that the coefficients $\alpha_i, i = 1, 2, \cdots$ are all regarded as known constants. The regression coefficient β can be derived from the likelihood function by minimizing the value of sum of least square.

$$\hat{\beta}(\alpha_i, i = 1, 2, \cdots) = \arg\inf_{\beta} \frac{1}{n} \sum_{t=1}^{n} [(Y_t - X_t\beta) - \sum_{k=1}^{\infty} \alpha_k (Y_{t-k} - X_{t-k}\beta)]^2 \tag{3.3}$$

Here, we denote $\hat{\beta}(\alpha_i, i = 1, 2, \cdots)$ as $\hat{\beta}$, then if we return to the estimation of α_i , the following results can be derive from minimizing the variance.

$$\hat{\alpha}^n = \arg\inf_{\alpha \in \Omega_K^p(A)} \frac{1}{n} \sum_{t=1}^{n} [(Y_t - X_t\hat{\beta}) - \sum_{k=1}^{\infty} \alpha_k (Y_{t-k} - X_{t-k}\hat{\beta})]^2 \tag{3.4}$$

In the framework of minimizing the least square estimation, the $\hat{\alpha}^n$ can be estimated after the estimation of β. However, this process is hard to handle with because the infinite AR error term is not easy to be estimated. $(Y_t - X_t\hat{\beta})$ is the estimated term, so the properties of $(Y_t - X_t\beta)$ will be useless in our estimation.

Thus, we change our framework into Bayesian one. From Bayesian framework, given a prior distribution of β, α can be estimated from the Bayesian formula. The remaining work for us is to maximize the posteriori likelihood function. Based on the distribution assumption of the AR error term, the following can be obtained

$$(\alpha^n, \beta, \sigma^2) = \arg \sup_{\alpha \in \Omega_K^p(A), \beta, \sigma > 0} L(\alpha, \beta | x, y) =$$

$$\arg \sup_{\alpha \in \Omega_K^p(A), \beta, \sigma > 0} \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{[(Y_t - X_t\beta) - \sum_{k=1}^{\infty} \alpha_k (Y_{t-k} - X_{t-k}\beta)]^2}{2\sigma^2}\} \tag{3.5}$$

We assume β and σ^2 have a joint distribution π, when given the sample X_t and Y_t. We denote the distribution as $\pi_{N,K}(\beta, \sigma^2 | x, y)$. So,

$$L(\alpha, \beta, \sigma^2 | x, y) \propto L(\alpha | \beta, \sigma^2, x, y) \pi_{N,K}(\beta, \sigma^2 | x, y) \tag{3.6}$$

Because the $\pi_{N,K}(\beta, \sigma^2 | x, y)$ can be given before we do the estimation. Different prior distributions contribute to difference results. So, besides the choices of prior distribution, the most important work for maximizing $L(\alpha, \beta, \sigma^2 | x, y)$ is to maximize $L(\alpha | \beta, \sigma^2, x, y)$. Thus, we derive from the similar results

as minimizing the least square estimation but applicable for consistency studies. We still denote the estimator as $\hat{\alpha}^n$ because the estimator from least square estimation will not be considered further.

$$\hat{\alpha}^n = \arg \inf_{\alpha \in \Omega_K^p(A)} L(\alpha|\beta, x, y)$$

$$\arg \inf_{\alpha \in \Omega_K^p(A)} \frac{1}{n} \sum_{t=1}^{n} [(Y_t - X_t\beta) - \sum_{k=1}^{\infty} \alpha_k (Y_{t-k} - X_{t-k}\beta)]^2 \tag{3.7}$$

4. Consistency Results for Estimator. The following condition provides the original models in 2.1 and 2.2 some basic assumptions. And the following theorems and lemmas are all based on the condition stated.

CONDITION 1. *Suppose that Y_t and X_t follows the regression model in 2.1 and the error term follows the AR model in 2.2, with $\phi \in \Omega^p$ and $\lambda_k \propto k^{\lambda}$, where $\lambda > 1/p$. Meanwhile, the coefficients ϕ satisfy the condition that $1 - \sum_{k=1}^{\infty} \phi_k z^k = 0$ has no solution in the unit circle.*

The main theorem in the paper is that the estimator $\hat{\alpha}^n$ of the constrained problem can approximate the real coefficient ϕ in a certain converge rate.

4.1. *Theorems and Lemmas .*

THEOREM 4.1. *If $\phi \in \Omega^p(A)$,*

$$|\hat{\alpha}^n - \phi|_p^p = O_p(n^{-\frac{p}{2}(\frac{q\tau+p}{(p-1)q\tau}+p^2)} + (K+1)^{-p\lambda}/[(p\omega - 1)\ln^{p\omega-1}(K+1)]) \tag{4.1}$$

where $1/p + 1/q = 1$, $p \geq 2$, $\tau < \min\{\gamma, \lambda p\}$, $\omega > 1/p$

In the following part, the denotation $\lesssim$ represents less than a certain value up to an absolute constant. And denote $Z_t(\alpha) = \sum_{k=1}^{\infty} \alpha_k (Y_{t-k} - X_{t-k}\beta)$. The following four lemmas are stated and will be proved in the latter part.

LEMMA 4.1. *If $\mu \in \Omega^p(A)$, then*

$$\alpha_k \lesssim \frac{k^{-(p\lambda+1)/p}}{\ln^{\omega}(1+k)}, \ \exists \ \omega > 1/p \tag{4.2}$$

Based on the denotation $Z_t(\alpha) = \sum_{k=1}^{\infty} \alpha_k (Y_{t-k} - X_{t-k}\beta)$, define the empirical loss function and population loss function as follows

$$M_n(\alpha) := \frac{1}{n} \sum_{t=1}^{n} [(Y_t - X_t\beta) - Z_t(\alpha)]^2 \tag{4.3}$$

$$M(\alpha) := \mathbb{E}Z_1^2(\phi - \alpha) \tag{4.4}$$

Denote $\phi^K \in \mathbb{R}^\infty$ such that $\phi^K = \phi I(i \leq K)$ where i represent the i-th position in the vector ϕ.

The following part of proof is based on the theorem about convergence rate.(van der Vaart and Wellner, 2000, Theorem 3.2.5).

And the Condition 1 is the basis of all the theorems and lemmas in the paper.

LEMMA 4.2. *Base on condition 1, for sufficiently small $\theta > 0$, $|\alpha - \phi^K|_p^p < \theta$*

$$M(\phi^K) - M(\alpha) \lesssim -|\alpha - \phi^K|_p^p \tag{4.5}$$

which means that as for the population loss function, if there is a α in the neighborhood of ϕ^k, the difference between the population loss functions of the two value can be controlled by the p sub square of the distance between α and ϕ^k.

LEMMA 4.3. *Base on condition 1, for $\forall$ n , and sufficient small δ, the difference between empirical and population loss function satisfies*

$$E \sup_{|\alpha - \phi^K|_p < \delta} |(M_n - M)(\phi^K) - (M_n - M)(\alpha)| \lesssim \frac{\psi_n(\delta)}{\sqrt{n}} \tag{4.6}$$

where the function ψ_n satisfies that $\delta \to \psi_n(\delta)/\delta^\gamma$ is decreasing for some $0 < \gamma < p$ and γ is independent with n.

REMARK 4.1. Base on condition 1, the $\psi_n(\delta)$ stated in lemma 4.1 is the same as the function stated in lemma 4.3. Then, we can give a sequence r_n that satisfies

$$r_n^p \psi_n(\frac{1}{r_n}) \leq \sqrt{n} \quad ,\forall \ n \tag{4.7}$$

LEMMA 4.4. *Base on condition 1, the target sequence $\hat{\alpha}^n$ converges to ϕ^K.*

THEOREM 4.2. *Base on condition 1, the denotations are just like in the lemma 4.1, 4.2, 4.3, 4.4 and notation 4.1. Then,*

$$r_n|\hat{\alpha}^n - \phi^K|_p = O_p(1) \tag{4.8}$$

where the first p means a kind of distance and the second means probability.

4.2. *Proof for the Lemmas.* Proof for Lemma 4.1

LEMMA. *If $\alpha \in \Omega^p(A)$, then*

$$\alpha_k \lesssim \frac{k^{-(p\lambda+1)/p}}{\ln^{\omega}(1+k)}, \ \exists \ \omega > 1/p \tag{4.9}$$

PROOF. Considering the meaning of $\lesssim$, there can be an absolute constant coefficient in the inequality. So, it is sufficient to prove the inequality if the following statement is true.

$$\sum_{k=1}^{\infty}(\frac{k^{-(p\lambda+1)/p}}{\ln^{\omega}(1+k)})^p \lambda_k^p \leq A^p \tag{4.10}$$

$\Leftrightarrow$

$$C\sum_{k=1}^{\infty}\frac{k^{-(p\lambda+1)}}{\ln^{p\omega}(1+k)}k^{p\lambda} \leq A^p$$

$\Leftrightarrow$

$$C\sum_{k=1}^{\infty}\frac{1}{k\ln^{p\omega}(1+k)} \leq A^p$$

$\Leftrightarrow$

$$\frac{1}{\ln^{p\omega}(2)} + \int_2^{\infty}\frac{dk}{k\ln^{p\omega}(1+k)} \leq \frac{1}{\ln^{p\omega}(2)} + \int_2^{\infty}\frac{dk}{k\ln^{p\omega}(k)} =$$

$$\frac{1}{\ln^{p\omega}(2)} + \frac{1}{p\omega-1}\frac{1}{\ln^{p\omega-1}(2)} < \infty$$

where the fixed absolute constant C works for adjusting the inequality to satisfy A^2.

The left side in the last inequality converges to the finite. Thus, the lemma is proved. □

Proof for lemma 4.2

LEMMA. *Base on condition 1, for sufficiently small $\theta > 0$, $|\alpha - \phi^K|_p^p < \theta$*

$$M(\phi^K) - M(\alpha) \lesssim -|\alpha - \phi^K|_p^p \tag{4.11}$$

PROOF. In (Sancetta, 2018), we know

$$M(\phi^K) - M(\alpha) \lesssim -|\alpha - \phi^K|_2^2 \lesssim -|\alpha - \phi^K|_p^p$$

where the second inequality is because the constraint of $|\alpha - \phi^K|_p^p < \theta$. □

Proof for lemma 4.3

LEMMA. *Base on condition 1, for $\forall$ n , and sufficient small δ, the difference between empirical and population loss function satisfies*

$$E \sup_{|\alpha-\phi^K|_p<\delta} |(M_n - M)(\phi^K) - (M_n - M)(\alpha)| \lesssim \frac{\psi_n(\delta)}{\sqrt{n}} \tag{4.12}$$

where the functions ψ_n satisfy that $\delta \to \psi_n(\delta)/\delta^\gamma$ is decreasing for some $0 < \gamma < p$ and γ is independent with n.

PROOF. Apply the denotation $Z_t(\alpha) = \sum_{k=1}^{\infty} \alpha_k(Y_{t-k} - X_{t-k}\beta)$, we can simplify the target formula.

$$(M_n - M)(\phi^K) - (M_n - M)(\alpha) =$$

$$\frac{1}{n}\sum_{t=1}^{n}\{2u_t Z_t(\phi^K - \alpha) + [Z_t^2(\phi - \phi^K) - Z_t^2(\phi - \alpha)] - [EZ_t^2(\phi - \phi^K) - EZ_t^2(\phi - \alpha)]\} \tag{4.13}$$

The combination of the second term and the last term above can be transformed into

$$|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(I - E)Z_t^2(\alpha - \phi^K)Z_t^2(2\phi - \phi^K - \alpha)| \lesssim \tag{4.14}$$

where we denote $\alpha - \phi^K = v$, $2\phi - \phi^K - \alpha = w$

$$E \sup_{\substack{w\in\Omega^p(4A),\\ v\in\Omega^p(2A),|v|_p\le\delta}} \sum_{l=1}^{+\infty}|v_l| \sum_{k=1}^{+\infty}|w_k||\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(I - E)(Y_{t-k} - X_{t-k}\beta)(Y_{t-l} - X_{t-l}\beta)|$$

Apply Lemma 3 and 4 in (Sancetta, 2018), we can obtain that

$$|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(I - E)(Y_{t-k} - X_{t-k}\beta)(Y_{t-l} - X_{t-l}\beta)| < \infty$$

There exits $r > 1$ such that the formula above can be written as

$$\lesssim \sum_{k>1}^{+\infty} E \sup_{v\in\Omega^p(2A),|v|_p\le\delta} \sum_{l=1}^{+\infty}|v_l||\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(I - E)(Y_{t-k} - X_{t-k}\beta)(Y_{t-l} - X_{t-l}\beta)|$$

$$\lesssim \sup_{k>0} k^{-r} E \sup_{v\in\Omega^p(2A),|v|_p\le\delta} \sum_{l=1}^{+\infty}|v_l||\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(I - E)(Y_{t-k} - X_{t-k}\beta)(Y_{t-l} - X_{t-l}\beta)|$$

where $\sum_{l=1}^{+\infty}$ can be split into two terms $\sum_{l=1}^{L}+\sum_{l=L+1}^{\infty}$. And the first term can be bounded by Cauchy-Schwarz inequality,

$$\sup_{k>0}\{\sup_{v\in\Omega^p(2A),|v|_p\leq\delta}\sum_{l=1}^{L}|v_l|^p\}^{\frac{1}{p}}$$

$$\{\sum_{l=1}^{L}E|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(I-E)(Y_{t-k}-X_{t-k}\beta)(Y_{t-l}-X_{t-l}\beta)|^q\}^{\frac{1}{q}}\lesssim\delta L^{\frac{1}{q}} \tag{4.15}$$

Similarly, the second term can also be bounded.

$$\sup_{k>0}\{\sup_{v\in\Omega^p(2A),|v|_p\leq\delta}\sum_{l=L+1}^{+\infty}|v_l|^p k^{\lambda p-\tau}\}^{\frac{1}{p}}$$

$$\{\sum_{l=L+1}^{\infty}k^{-(\lambda p-\tau)}E|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(I-E)(Y_{t-k}-X_{t-k}\beta)(Y_{t-l}-X_{t-l}\beta)|^q\}^{\frac{1}{q}} \tag{4.16}$$

where τ is sufficiently small for $\lambda p-\tau>0$

Thus, the term with $1/q$ index is bounded. As for the remaining term,

$$\sup_{v\in\Omega^p(2A),|v|_p\leq\delta}\sum_{l=L+1}^{+\infty}|v_l|^p k^{\lambda p-\tau}\lesssim L^{\lambda p-\tau}\lambda_L^{-p}$$

Thus, 4.14 is bounded by $\delta L^{\frac{1}{q}}+L^{\frac{\lambda p-\tau}{p}}\lambda_L^{-1}\lesssim\delta L^{\frac{1}{q}}+L^{-\frac{\tau}{p}}$ where $\lambda p-\tau>0$

Considering the L can be chosen by us. Let $L=\delta^{-\frac{pq}{q\tau+p}}$

Thus, 4.14 is bounded by $\delta^{\frac{q\tau}{q\tau+p}}$, and it may be multiplied by a constant.

And use the same way of proving,

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}2u_tZ_t(\phi^K-\alpha)\lesssim\delta^{\frac{q\tau}{q\tau+p}}$$

Here, we choose $\psi_n(\delta)=\delta^{\frac{q\tau}{q\tau+p}}$. And let $\frac{q\tau}{q\tau+p}-\gamma<0$, then $\psi_n(\delta)/\delta^\gamma$ decreases. Thus we make $\tau<\min\{\gamma,\lambda p\}$. So far, the lemma is proven. □

Proof for lemma 4.4

LEMMA. *Base on condition 1, the target sequence $\hat{\alpha}^n$ converges to ϕ^K.*

PROOF. This lemma is a direct result from the absolutely summable coefficients. □

4.3. *Proof for the Theorems.* Proof for Theorem 4.2

THEOREM. *Base on condition 1, the denotations are just like in the lemma 4.1, 4.2, 4.3, 4.4 and notation 4.1. Then,*

$$r_n|\hat{\alpha}^n - \phi^K|_p = O_p(1) \tag{4.17}$$

where the first p means a kind of distance and the second means probability.

PROOF. According to the definition of loss function, we need to minimize the loss function. Denote $\hat{\alpha}^n$ which minimizes the function $\alpha \to M_n(\alpha)$.For each fixed n which is related to r_n, the parameter space $\mathcal{A}_n$ where ϕ^K is excluded, can be partitioned into several parts:

$$P_{i,n} = \{\alpha \in \mathbb{R}^\infty : p^{i-1} < r_n|\alpha - \phi^K|_p \leq p^i\}, i \in \mathbb{Z} \tag{4.18}$$

$$\bigcup_{i\in\mathbb{Z}} P_{i,n} = \mathcal{A}\backslash\{\phi^K\} \tag{4.19}$$

Focus on the result of the proof, $r_n|\hat{\alpha}^n - \phi^K|_p = O_p(1)$. If we verify that

$$\lim_{M\to+\infty} P(r_n|\hat{\alpha}^n - \phi^K|_p > p^M) = 0 \tag{4.20}$$

then the theorem is proved.

Here, if $r_n|\hat{\alpha}^n - \phi^K|_p > p^M$, and for a sufficient small positive ρ, $\{\hat{\alpha}^n \in \mathbb{R}^\infty | r_n|\hat{\alpha}^n - \phi^K|_p > p^M\}$ can be divided into two parts. The partition $P_{i,n}$ is used to give the first part. Here,

$$\bigcup_{i>M,\ p^i\leq\rho r_n} \{\hat{\alpha}^n \in \mathbb{R}^\infty || \hat{\alpha}^n - \phi^K|_p > \frac{\rho}{p^{i-M}}\} \subseteq$$

$$\bigcup_{i>M,\ p^i\leq\rho r_n} \{\sup_{\alpha\in P_{i,n}} (M_n(\phi^K) - M_n(\alpha)) \geq 0\}$$

The remaining part satisfies $p^M > \rho r_n$. Thus,

$$\{\hat{\alpha}^n \in \mathbb{R}^\infty | r_n|\hat{\alpha}^n - \phi^K|_p > p^M\} \subseteq \{\hat{\alpha}^n \in \mathbb{R}^\infty || \hat{\alpha}^n - \phi^K|_p \geq \frac{\rho}{p}\}$$

Based on the two parts, the following inequality can be obtained

$$P(r_n|\hat{\alpha}^n - \phi^K|_p > p^M) \leq P(\frac{\rho}{p} \leq |\hat{\alpha}^n - \phi^K|_p) +$$

$$\sum_{i>M,\ p^i\leq\rho r_n} P(\sup_{\alpha\in P_{i,n}} (M_n(\phi^K) - M_n(\alpha)) \geq 0) \tag{4.21}$$

Apply the lemma 4.4, the sequence $\hat{\alpha}^n$ converges to ϕ^K, thus,

$$\lim_{n\to\infty} P(|\hat{\alpha}^n - \phi^K|_p \geq \frac{\rho}{p}) = 0,\ \forall\ \rho > 0 \tag{4.22}$$

Choose ρ and $|\alpha - \phi^K|_p^p \leq \rho$ satisfy the lemma 4.2 , and $\rho < \delta$ in the lemma 4.3. Fix ρ and consider the result of the lemma 4.2 for each $i > M$, and for every $\alpha \in P_{i,n}$

$$M(\phi^K) - M(\alpha) \leq -|\alpha - \phi^K|_p^p \leq -\frac{p^{p(i-1)}}{r_n^p} \tag{4.23}$$

Thus,

$$\sum_{i>M,\ p^i\leq\rho r_n} P(\sup_{\alpha\in P_{i,n}} (M_n(\phi^K) - M_n(\alpha)) \geq 0) \leq$$

$$\sum_{i>M,\ p^i\leq\rho r_n} P(\sup_{\alpha\in P_{i,n}} ((M_n - M)(\phi^K) - (M_n - M)(\alpha)) \geq M(\alpha) - M(\phi^K) \geq \frac{p^{p(i-1)}}{r_n^p}) \tag{4.24}$$

Apply Markov's inequality, the definition of r_n, and the property that $\psi_n(\delta)/\delta^\gamma$ decreases for some $0 < \gamma < p$. Use notation 4.1

$$RHS\ of\ 4.24 \lesssim \sum_{i>M,\ p^i\leq\rho r_n} \frac{\psi_n(p^i/r_n) r_n^p}{\sqrt{n} p^{pi}} \lesssim \sum_{i>M} (p^i)^{\gamma-p} \to_{M\to\infty} 0 \tag{4.25}$$

Thus,

$$\lim_{M\to+\infty} P(r_n|\hat{\alpha}^n - \phi^K|_p > p^M) = 0$$

The theorem 4.2 is proved. □

Then let us return to the theorem 4.1 and show how to prove it.
Proof for Theorem 4.1

THEOREM. *If* $\phi \in \Omega^p(A)$,

$$|\hat{\alpha}^n - \phi|_p^p = O_p(n^{-\frac{p}{2}(\frac{q\tau+p}{(p-1)q\tau}+p^2)} + (K+1)^{-p\lambda}/[(p\omega-1)\ln^{p\omega-1}(K+1)]) \tag{4.26}$$

where $1/p + 1/q = 1$, $p \geq 2$, $\tau < min\{\gamma, \lambda p\}$, $\omega > 1/p$

PROOF. If $\phi \in \Omega^p(A)$,

$$|\hat{\alpha}^n - \phi|_p^p = |\hat{\alpha}^n - \phi^K|_p^p + |\phi^K - \phi|_p^p \tag{4.27}$$

where $|\phi^K - \phi|_p$ is the remaining term when the first K entries of ϕ are deleted. Apply lemma 4.1, because $\phi \in \Omega^p(A)$,

$$|\phi^K - \phi|_p^p = \sum_{j=K+1}^{\infty} |\phi_j|^p \lesssim \sum_{j=K+1}^{\infty} (\frac{j^{-(p\lambda+1)/p}}{\ln^{\omega}(1+j)})^p \lesssim$$
$$\int_{K+1}^{\infty} \frac{j^{-(p\lambda+1)}}{\ln^{p\omega}(j)} dj \leq (K+1)^{-p\lambda} \int_{K+1}^{\infty} \frac{1}{j * \ln^{p\omega}(j)} dj =$$
$$(K+1)^{-p\lambda}/[(p\omega - 1)\ln^{p\omega-1}(K+1)]$$

Thus,

$$|\phi^K - \phi|_p^p \lesssim O_p((K+1)^{-p\lambda}/[(p\omega - 1)\ln^{p\omega-1}(K+1)]) \tag{4.28}$$

As for $|\hat{\alpha}^n - \phi^K|_p^p$, we can apply the theorem 4.2. If we can verify r_n, and

$$|\hat{\alpha}^n - \phi^K|_p^p = O_p(r_n^{-p}) \tag{4.29}$$

According to the notation 4.1,

$$r_n^p \psi_n(\frac{1}{r_n}) \leq \sqrt{n} \quad , \quad \psi_n(\delta) = \delta^{\frac{q\tau}{q\tau+p}} \tag{4.30}$$

which means

$$r_n \leq n^{\frac{1}{2}(\frac{q\tau+p}{(p-1)q\tau} + p^2)}$$

Thus,

$$|\hat{\alpha}^n - \phi^K|_p^p = O_p(n^{-\frac{p}{2}(\frac{q\tau+p}{(p-1)q\tau} + p^2)})$$

Thus,

$$|\hat{\alpha}^n - \phi|_p^p = O_p(n^{-\frac{p}{2}(\frac{q\tau+p}{(p-1)q\tau} + p^2)} + (K+1)^{-p\lambda}/[(p\omega - 1)\ln^{p\omega-1}(K+1)]) \tag{4.31}$$

And we finish proving the theorem 4.1.

□

5. Application. The parameter B in ellipsoid is chosen to minimize the prediction error in cross-validation estimation, (Burman and Nolan, 1992). In practice, we often minimize the prediction error with penalized terms. One may refer to Hastie et al. (2009) for more details about the penalized terms and refer to Sancetta (2018) for the choice of B. Here, we apply the criterion shown in Section 2.2 in Sancetta (2018).

We conduct an empirical application to provide insights into the performance and properties of our model when fitting data. Suppose we have data about the simulation of an AR(T) model,

$$\phi_k = \phi_0 \frac{k^{-1/p}}{\sum_{k=1}^{T} k^{-1/p}} \tag{5.1}$$

where, ϕ_k's are the coefficients of AR(T) model and p chosen in 2.4. We notice that there will always be an ellipsoid if the parameter T is finite. It is clear that we choose the coefficients like 5.1 to satisfy the requirement 1. When $0 < \phi_0 < 1$, the requirement 1 is easy to meet. And the coefficients are less when k grows. Besides the ar term, we consider the regression term. Then, we use linear regression model with only one variable, and interception, slope equal to 0.2, 2, separately. The explanatory variable is uniform distribution which varies from 0 to 20. The error term is the ar term given above. And we set $\phi_0 = 0.90$, for instance. We choose the true value of T to be 100. Then we choose $T = 100, 1000, 10000, 100000$ separately to be the parameter of our estimated model to see if our model is consistent.

Then, we perform the maximum likelihood estimation to predict the value of ϕ_0. As we set, the original value of T is 100. With the increase of T, the length of lag terms increases. And the constraint of ellipsoid will work to keep our model consistent. When considering the choice of A, the method in the section 2.2 of Sancetta (2018) can be applied.

From Table 1, we can observe the consistence.

TABLE 1
estimated $\hat{\phi}_0$ for chosen models

T	100	1000	10000	1000000
$\hat{\phi}_0$	0.99	0.99	0.99	0.99

From 1, under the constraint of ellipsoid, the performance of our model is robust and consistent with the increase of the number of coefficients.

References.

Beach, C. M., MacKinnon, J. G., 1978. A maximum likelihood procedure for regression with autocorrelated errors. Econometrica: Journal of the Econometric Society, 51–58

Burman, P., Nolan, D., 1992. Datadependent estimation of prediction functions. Journal of Time Series Analysis 13 (3), 189–207

Bhlmann, P., 1995. Moving-average representation of autoregressive approximations. Stochastic processes and their applications 60 (2), 331–342

Bhlmann, P., 1997. Sieve bootstrap for time series. Bernoulli 3 (2), 123–148

Chib, S., 1993. Bayes regression with autoregressive errors: A gibbs sampling approach. Journal of Econometrics 58 (3), 275–294

Cochrane, D., Orcutt, G. H., 1949. Application of least squares regression to relationships containing auto-correlated error terms. Journal of the American statistical association 44 (245), 32–61

DeGroot, M. H., 2005. Optimal statistical decisions. Vol. 82. John Wiley & Sons.

Doucet, A., Godsill, S. J., Robert, C. P., 2002. Marginal maximum a posteriori estimation using markov chain monte carlo. Statistics and Computing 12 (1), 77–84.

Gauvain, J.-L., Lee, C.-H., 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE transactions on speech and audio processing 2 (2), 291–298.

Greig, D. M., Porteous, B. T., Seheult, A. H., 1989. Exact maximum a posteriori estimation for binary images. Journal of the Royal Statistical Society. Series B (Methodological), 271–279.

Hastie, T., Tibshirani, R., Friedman, J., 2009. Unsupervised learning. In: The elements of statistical learning. Springer, pp. 485–585.

Leggetter, C. J., Woodland, P. C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Computer speech & language 9 (2), 171–185.

Lokshin, M., Sajaia, Z., et al., 2004. Maximum likelihood estimation of endogenous switching regression models. Stata Journal 4, 282–289.

Nerurkar, E. D., Roumeliotis, S. I., Martinelli, A., 2009. Distributed maximum a posteriori estimation for multi-robot cooperative localization. In: Robotics and Automation, 2009. ICRA'09. IEEE International Conference on. IEEE, pp. 1402–1409.

Pierce, D. A., 1971. Least squares estimation in the regression model with autoregressive-moving average errors. Biometrika 58 (2), 299–312 1464–3510.

Sacks, J., Ylvisaker, D., 1966. Designs for regression problems with correlated errors. The Annals of Mathematical Statistics 37 (1), 66–89.

Sancetta, A., 2018. Consistency results for stationary autoregressive processes with constrained coefficients. IEEE Transactions on Information Theory.

Schafer, D., 2002. Strongly consistent online forecasting of centered gaussian processes. IEEE Transactions on Information Theory 48 (3), 791–799.

Department of Mathematical Science
Tsinghua University
Beijing 100084, China
E-mail: zrz6787@gmail.com

Center for Statistical Science and
Department of Industrial Engineering
Tsinghua University
Beijing 100084, China
E-mail: malidong@tsinghua.edu.cn