Negative Moment Bounds for Stochastic Regression Models with Deterministic Trends and Their Applications to Prediction Problems

Chien-Ming Chi
University of Southern California

Ching-Kang Ing
National Tsing Hua University

Shu-Hui Yu
National University of Kaohsiung

November 3, 2020

Abstract

We establish negative moment bounds for the minimum eigenvalue of the normalized Fisher information matrix in a stochastic regression model with a deterministic time trend. This result enables us to develop an asymptotic expression for the mean squared prediction error (MSPE) of the least squares predictor in the aforementioned model. Our asymptotic expression not only helps better understand how the MSPE is affected by the deterministic and random components, but also inspires an intriguing proof of the formula for the sum of elements in the inverse of the Cauchy/Hilbert matrix from a prediction perspective.

Keywords: Cauchy matrix, Hilbert matrix, mean squared prediction error, minimum eigenvalue, negative moment bound, stochastic regression model

1 Introduction

The stochastic regression model is one of the most widely used statistical models due to its broad applications in engineering, economics, medicine, and many other scientific fields. In their seminal paper, Lai and Wei (1982) laid the theoretical foundations of parameter estimation in such a model. In particular, they proposed a set of weakest possible conditions under which the linear least squares estimate achieves strong consistency. Lai and Wei's (1982) paper inspired a great deal of exciting work, bringing insights into prediction, model selection, non-linear estimation, stochastic approximation, and adaptive control; see, e.g., Chen and Guo (1986), Lai and Wei (1986), Wei (1987, 1992), Lai (1994), Lai and Lee (1997), Chen *et al.* (1999), and Gerencsér *et al.* (2009).

One of the most important purposes of statistical modeling is to predict future values. The performance of a prediction method is usually evaluated by two different measures: the accumulated prediction error (APE) and the mean squared prediction error (MSPE). Model selection based on these two types of errors also attracted a lot of attention from researchers and practitioners. Wei (1987) provided asymptotic expressions for the APEs of the least squares predictors in stochastic regression models. Model selection based on the APE was explored by Rissanen (1986), Wax (1988), Hannan *et al.* (1989), Hemerly and Davis (1989), Wei (1992), Speed and Yu (1994), West (1996), Lai and Lee (1997), Ing (2004, 2007), Ing *et al.* (2009), and Ing and Yang (2014). Asymptotic expressions for the MSPEs of least squares predictors were derived in a variety of time series models; see, e.g., Fuller and Hasza (1981), Kunitomo and Yamamoto (1985), Gerencsér (1992), Ing (2003), Ing *et al.* (2009), Chan and Ing (2011), and Chan *et al.* (2013). There are also quite a few model selection methods proposed based on minimizing the MSPE; see Shibata (1980), Bhansali (1996), Lee and Karagrigoriou (2001), Ing and Wei (2005), Ing *et al.* (2012), and Hsu *et al.* (2019).

Since many time series data exhibit polynomial or other deterministic time trends, parameter estimation and hypothesis testing in time series models with drifts have been considered by a number of authors; see, e.g., Chan (1989), Hamilton (1994), and Stock (1994). On the other hand, most existing studies on the MSPE have focused on the case where the underlying time series model has a constant mean. Although Ing (2003) derived an asymptotic expression for the MSPE of the least squares predictor in an autoregressive (AR) model around a polynomial trend, it seems difficult to apply his result to more general time series models. In addition, his derivation heavily relied on a negative moment bound for the minimum eigenvalue of the normalized Fisher information matrix of a non-constant mean, whose proof, however, was not rigorously given. In this paper, we fill this gap by investigating the MSPEs of the least squares predictors in autoregressive exogenous (ARX) models (an important class of stochastic regression models) with deterministic trends satisfying general conditions. We first establish negative moment bounds for the minimum eigenvalue of the normalized Fisher information matrix, $\hat{\boldsymbol{R}}_n$, associated with this model in a rigorous manner. With the help of these bounds, we provide an asymptotic expression for the MSPE of the least squares predictor, which is the sum of two terms accounting for the variations due to estimating the time trend and ARX components, respectively. This result helps us better understand how the MSPE is affected by the model's deterministic and random components.

Our asymptotic expression shows that the MSPE due to estimating the polynomial time trend is related to the sum of elements in the inverse of the Hilbert matrix, which, in turn, is a special case of the symmetric Cauchy matrix. The formula for the sum of elements in the latter matrix's inverse was given by Schechter (1959) through Lagrange's interpolation method. The connection between the MSPE and Cauchy/Hilbert matrix triggers us to ask if there is an alternative proof of the formula from a prediction perspective. By establishing an intriguing link between the MSPE and APE, we show that the answer to this question is affirmative.

The rest of the paper is organized as follows. In Section 2, we establish negative moment bounds for the minimum eigenvalue of a matrix associated with $\hat{\boldsymbol{R}}_n$. In Section 3.1, asymptotic expressions for the MSPEs of the least squares predictors in ARX models with general time trends are given. We further illustrate the results using the polynomial and periodic time trends. In Section 3.2, a statistical proof of the formula for the sum of elements in the inverse of the Cauchy matrix is provided. We conclude in Section 4. All proofs of the theorems in Sections 2 and 3.1 and other technical details are relegated to the Appendix.

2 Negative Moment Bounds

Let k and m be positive integers. We start by considering a km-dimensional time series,

$$\boldsymbol{Y}_t = \sum_{j=0}^{\infty} C_j \boldsymbol{\varepsilon}_{t,j}, \tag{2.1}$$

where $\boldsymbol{\varepsilon}_{t,j} = (\boldsymbol{\delta}_{t-jk}^{\top}, \ldots, \boldsymbol{\delta}_{t-(j+1)k+1}^{\top})^{\top}$, $\{\boldsymbol{\delta}_t\}$ is a sequence of m-dimensional independent random vectors satisfying $\mathbf{E}(\boldsymbol{\delta}_t) = \mathbf{0}$ and $\mathbf{E}(\boldsymbol{\delta}_t \boldsymbol{\delta}_t^{\top}) = \boldsymbol{\Sigma} > 0$, C_js are $km \times km$ coefficient matrices, C_0 is invertible, and

$$\sum_{j=0}^{\infty} \|C_j\|_F^2 < \infty. \tag{2.2}$$

Here, $\|A\|_F$ denotes the Frobenius norm of matrix A. Many time series regression models have explanatory vectors satisfying (2.1). Here, we give two examples.

Example 2.1 *Let* $\boldsymbol{z}_t = \sum_{j=0}^{\infty} D_j \boldsymbol{\delta}_{t-j}$ *be an* m*-dimensional stationary time series, where* $\sum_{j=0}^{\infty} \|D_j\|_F^2 < \infty$, D_0 *is invertible, and* $\{\boldsymbol{\delta}_t\}$ *is defined as in model* (2.1). *Then,*

$$\begin{pmatrix} \boldsymbol{z}_t \\ \vdots \\ \boldsymbol{z}_{t-k+1} \end{pmatrix} = \sum_{j=0}^{\infty} \begin{pmatrix} D_{jk} & \cdots & D_{(j+1)k-1} \\ \vdots & & \vdots \\ D_{(j-1)k+1} & \cdots & D_{jk} \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta}_{t-jk} \\ \vdots \\ \boldsymbol{\delta}_{t-(j+1)k+1} \end{pmatrix}, \tag{2.3}$$

where $D_l = \mathbf{0}$ if $l < 0$. Hence (2.1) *holds with $\mathbf{Y}_t = (\boldsymbol{z}_t^\top, \cdots, \boldsymbol{z}_{t-k+1}^\top)^\top$,*

$$C_j = \begin{pmatrix} D_{jk} & \cdots & D_{(j+1)k-1} \\ \vdots & & \vdots \\ D_{(j-1)k+1} & \cdots & D_{jk} \end{pmatrix},$$

and $\boldsymbol{\varepsilon}_{t,j} = (\boldsymbol{\delta}_{t-jk}^\top, \cdots, \boldsymbol{\delta}_{t-(j+1)k+1}^\top)^\top$. One may use the following vector AR model to make prediction,

$$\boldsymbol{z}_{t+1} = \sum_{j=1}^{k} \mathbf{\Theta}_j \boldsymbol{z}_{t+1-j} + \boldsymbol{\epsilon}_{t+1}, \tag{2.4}$$

where $\mathbf{\Theta}_j$s are $m \times m$ coefficient matrices and $\boldsymbol{\epsilon}_{t+1}$ is the model error, which can be serially correlated if (2.4) *is misspecified. It is clear that the explanatory vector of model* (2.4) *is given on the left-hand side of* (2.3).

Example 2.2 *Consider an autoregressive exogenous (ARX) model,*

$$v_t = \sum_{j=1}^{k_0} a_j v_{t-j} + \sum_{l=1}^{d} \sum_{j=1}^{k_l} \theta_j(l) z_{t-j}(l) + \varepsilon_t, \tag{2.5}$$

where $d, k_0, \ldots, k_d$ are positive integers, a_j and $\theta_j(l)$ are unknown coefficients,

$$1 - a_1 z - \cdots - a_{k_0} z^{k_0} \neq 0, \; |z| \leq 1, \tag{2.6}$$

$(z_{t-1}(l), \ldots, z_{t-k_l+1}(l))^\top, l = 1, \ldots, d,$ are exogenous variables admitting the following

MA(∞) representation,

$$z_t(l) = \sum_{j=0}^{\infty} b_j(l)\varepsilon_{t-j}(l), \tag{2.7}$$

with $b_0(l) = 1$ *and* $\sum_{j=0}^{\infty} b_j^2(l) < \infty$*, and* $\boldsymbol{\delta}_t = (\varepsilon_t, \varepsilon_t(1), \ldots, \varepsilon_t(d))^\top, t = 1, \ldots, n$*, are independent noises satisfying* $\mathbf{E}(\boldsymbol{\delta}_t) = \mathbf{0}$ *and* $(\sigma_{ij})_{1 \leq i,j \leq d+1} = \mathbf{E}(\boldsymbol{\delta}_t \boldsymbol{\delta}_t^\top) > 0$*. By* (2.6) *and* (2.7)*, there exist* $\boldsymbol{\eta}_j, j \geq 0$*, with* $\boldsymbol{\eta}_0 = (1, 0, \ldots, 0)^\top$ *and* $\sum_{j=1}^{\infty} \|\boldsymbol{\eta}_j\|^2 < \infty$*, such that*

$$v_t = \sum_{j=0}^{\infty} \boldsymbol{\eta}_j^\top \boldsymbol{\delta}_{t-j}, \tag{2.8}$$

noting that $\|\cdot\|$ *denotes the Euclidean norm. Let* $\bar{k} = \max\{k_0, \ldots, k_d\}$*. Then*

$$\boldsymbol{Y}_t = (v_t, z_t(1), \ldots, z_t(d), \ldots, v_{t-\bar{k}+1}, z_{t-\bar{k}+1}(1), \ldots, z_{t-\bar{k}+1}(d))^\top \tag{2.9}$$

can be viewed as an explanatory vector of model (2.5) *containing possibly redundant components. It follows from* (2.7) *and* (2.8) *that*

$$\boldsymbol{Y}_t = \sum_{j=0}^{\infty} C_j \boldsymbol{\varepsilon}_{t,j},$$

where $\boldsymbol{\varepsilon}_{t,j} = (\boldsymbol{\delta}_{t-j\bar{k}}^\top, \ldots, \boldsymbol{\delta}_{t-(j+1)\bar{k}+1}^\top)^\top$ *and*

$$C_j = \begin{pmatrix} C_{j,11} & \cdots & C_{j,1\bar{k}} \\ \vdots & & \vdots \\ C_{j,\bar{k}1} & \cdots & C_{j,\bar{k}\bar{k}} \end{pmatrix},$$

in which

$$C_{j,tl} = \begin{pmatrix} \eta_{j\bar{k}-t+l,1} & \eta_{j\bar{k}-t+l,2} & \cdots & \cdots & \eta_{j\bar{k}-t+l,d+1} \\ 0 & b_{j\bar{k}-t+l}(1) & 0 & \cdots & 0 \\ \vdots & 0 & b_{j\bar{k}-t+l}(2) & 0 & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & b_{j\bar{k}-t+l}(d) \end{pmatrix}$$

is a $(d+1)\times(d+1)$ *matrix with* $(\eta_{j\bar{k}-t+l,1}, \ldots, \eta_{j\bar{k}-t+l,d+1})^\top = \boldsymbol{\eta}_{j\bar{k}-t+l}$ *and* $\eta_{h,l_1} = b_h(l_2) = 0$ *if* $h < 0$*. Since* $\sum_{j=0}^{\infty} \|C_j\|_F^2 < \infty$ *and* C_0 *is invertible due to* $C_{0,tt} = I_{d+1}$ *(the* $(d+1)$*-dimensional identity matrix) and* $C_{0,tl} = \mathbf{0}$ *if* $t > l$*, we conclude that the* $\mathbf{Y}_t$ *in* (2.9) *fulfills* (2.1) *with* $k = \bar{k}$ *and* $m = d+1$*.*

Assuming that (2.1) holds and there exist $\delta, M, \alpha > 0$ such that for any $0 < w-u \leq \delta$,

$$\sup_{-\infty<t<\infty} \sup_{\|\boldsymbol{\nu}\|=1} P(u < \boldsymbol{\nu}^\top \boldsymbol{\delta}_t \leq w) \leq M(w-u)^\alpha, \tag{2.10}$$

Findley and Wei (2002) showed that for any $q \geq 1$,

$$\mathbf{E}\left(\lambda_{\min}^{-q}(n^{-1}\sum_{t=1}^{n} \mathbf{Y}_t \mathbf{Y}_t^\top)\right) = O(1), \tag{2.11}$$

where n is the sample size and $\lambda_{\min}(A)$ denotes the minimum eigenvalue of matrix A. With the help of (2.11), they presented the first mathematically complete derivation of an analogous property of AIC for comparing vector autoregressions fit to weakly stationary series. Using (2.11) and an argument in Findley and Wei (2002), one can also obtain an asymptotic expression for the MSPE of the least squares predictors in model (2.4) (or (2.5)) in terms of the sample size, the variance of the model error, and the number of the estimated parameters.

When a deterministic trend (containing p variables with $p \geq 1$) is taken into account, a natural generalization of (2.11) is

$$\mathbf{E}\left(\lambda_{\min}^{-q}(n^{-1}\sum_{t=1}^{n}\boldsymbol{\omega}_t\boldsymbol{\omega}_t^{\top})\right) = O(1), \tag{2.12}$$

where $\boldsymbol{\omega}_t = (\boldsymbol{x}_t^{(n)^{\top}}, \boldsymbol{Y}_t^{\top})^{\top}$ and $\boldsymbol{x}_t^{(n)} \in R^p$, possibly depending on n, is the normalized time trend variables satisfying

$$\sup_{1\leq t\leq n}\|\boldsymbol{x}_t^{(n)}\| < M_1, \tag{2.13}$$

for some positive constant M_1. One would expect that (2.12) holds under the additional assumption,

$$\liminf_{n\to\infty}\lambda_{\min}(n^{-1}\sum_{t=1}^{n}\boldsymbol{x}_t^{(n)}\boldsymbol{x}_t^{(n)^{\top}}) > 0, \tag{2.14}$$

which is commonly made on the fixed design matrix. The proof of (2.12), however, is far from being trivial. The main reason is that the proof of (2.11) is built on the property that for any $\boldsymbol{a} \in R^{km}$ with $\|\boldsymbol{a}\| = 1$, the conditional distribution of $(\boldsymbol{a}^\top \boldsymbol{Y}_t)^2$ given information up to time $t - l$ is sufficiently smooth at the origin as long as l is large enough. This property is ensured by (2.10), but is no longer valid when $\boldsymbol{Y}_t$ is replaced by $\boldsymbol{\omega}_t$. With the appearance of $\boldsymbol{x}_t^{(n)}$, it is easy to find a unit vector $\boldsymbol{a} \in R^{km+p}$ such that $\boldsymbol{a}^\top \boldsymbol{\omega}_t = 0$. In Lemma A.2, we provide a characterization of (2.14). This characterization is not only of independent interest, but it also inspires a proof strategy to bypass the above difficulty. The main result of this section is given in the following theorem.

Theorem 2.1 *Assume* (2.1)*,* (2.2)*,* (2.10)*,* (2.13)*,* (2.14)*, and*

$$\sup_{-\infty<t<\infty} \max_{1\le i\le m} \mathbf{E}|\delta_{t,i}|^{2\gamma} < \infty, \tag{2.15}$$

where $\gamma > 1$ *and* $(\delta_{t,1}, \ldots, \delta_{t,m})^\top = \boldsymbol{\delta}_t$*. Then, for* $0 < q < \gamma$*,* (2.12) *follows.*

3 Applications

3.1 Mean squared prediction errors

In this section, we focus on the ARX model around a deterministic time trend,

$$y_t = \sum_{j=1}^{p} \beta_j s_{t,j} + \sum_{j=1}^{k_0} a_j y_{t-j} + \sum_{l=1}^{d} \sum_{j=1}^{k_l} \theta_j(l) z_{t-j}(l) + \varepsilon_t \tag{3.1}$$

where p, d, and $k_0, \ldots, k_d$ are positive integers, β_j, a_j, and $\theta_j(l)$ are unknown coefficients, with a_j satisfying (2.6), $\boldsymbol{z}_{t-1}(l) = (z_{t-1}(l), \ldots, z_{t-k_l}(l))^\top, 1 \leq l \leq d$, are exogenous variables admitting the MA(∞) representations described in (2.7), $\boldsymbol{s}_t = (s_{t,1}, \ldots, s_{t,p})^\top$ are deterministic variables, and $\boldsymbol{\delta}_t = (\varepsilon_t, \varepsilon_t(1), \ldots, \varepsilon_t(d))^\top$ are the same as the one in Example 2.2. Let $\boldsymbol{P}_t = (\boldsymbol{s}_t^\top, \boldsymbol{y}_{t-1}^\top, \boldsymbol{z}_{t-1}^\top(1), \ldots, \boldsymbol{z}_{t-1}^\top(d))^\top$, where $\boldsymbol{y}_t = (y_t, \ldots, y_{t-k_0+1})^\top$. Having observed $y_1, \ldots, y_n$ and $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_{n+1}$, we are interested in predicting y_{n+1} using the least squares predictor,

$$\hat{y}_{n+1} = \boldsymbol{P}_{n+1}^\top (\sum_{t=1}^{n} \boldsymbol{P}_t \boldsymbol{P}_t^\top)^{-1} \sum_{t=1}^{n} \boldsymbol{P}_t y_t, \tag{3.2}$$

provided the inverse of $\sum_{t=1}^{n} \boldsymbol{P}_t \boldsymbol{P}_t^\top$ exists.

To analyze the MSPE, $\mathbf{E}(y_{n+1} - \hat{y}_{n+1})^2$, of $\hat{y}_{n+1}$, we impose the following conditions on the deterministic terms $\boldsymbol{s}_t$: there exists a $p \times p$ matrix $\boldsymbol{D}$ such that for any t,

$$\boldsymbol{s}_{t-1} = \boldsymbol{D} \boldsymbol{s}_t, \tag{3.3}$$

and

$$I_p - \sum_{j=1}^{k_0} a_j \boldsymbol{D}^j \text{ is invertible.} \tag{3.4}$$

By (3.3) and (3.4), it can be shown that

$$y_t = {\boldsymbol{\beta}^*}^\top \boldsymbol{s_t} + v_t, \tag{3.5}$$

where v_t is defined in (2.8) and $\boldsymbol{\beta}^{*\top} = \boldsymbol{\beta}^\top(I_p - \sum_{j=1}^{k_0} a_j \boldsymbol{D}^j)^{-1}$ with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$. Many commonly used deterministic trends fulfill (3.3) and (3.4). For example, in the case of the polynomial trend,

$$\boldsymbol{s}_t = (1, t, \ldots, t^{p-1})^\top, p \geq 1, \tag{3.6}$$

we have $\boldsymbol{D} = (D_{ij})_{1 \leq i,j \leq p}$, where $D_{ij} = 0$ if $1 \leq i < j \leq p$, and $C_{i-j}^{i-1}(-1)^{i-j}$ if $1 \leq j \leq i \leq p$, where $C_{i-j}^{i-1} = (i-1)!/[(i-j)!(j-1)!]$. Since $\boldsymbol{D}^j, j \geq 1$, are lower triangular matrices with diagonal entries 1, (3.4) holds when (2.6) is assumed. For the periodic trend,

$$\boldsymbol{s}_t = (1, \sin\nu_1 t, \cos\nu_1 t \ldots, \sin\nu_h t, \cos\nu_h t)^\top, \tag{3.7}$$

where $h \geq 1$ and $0 < \nu_1 < \cdots < \nu_h < \pi$, we have $\boldsymbol{D} = \mathrm{Diag}(1, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_h)$, where

$$\boldsymbol{\nu}_j = \begin{pmatrix} \cos\nu_j & -\sin\nu_j \\ \sin\nu_j & \cos\nu_j \end{pmatrix}.$$

Also, (3.4) follows from (2.6). By (3.3) and (3.5), the trend in $\boldsymbol{y}_{t-1}$ can be removed through a linear transformation of $\boldsymbol{P}_t$,

$$\boldsymbol{y}_{t-1} - G\boldsymbol{s}_t = \boldsymbol{v}_{t-1} = (v_{t-1}, \ldots, v_{t-k_0})^\top, \tag{3.8}$$

where $G^\top = (\boldsymbol{D}^\top\boldsymbol{\beta}^* \cdots \boldsymbol{D}^{k_0^\top}\boldsymbol{\beta}^*)$. Suppose that there exists a $p \times p$ non-random matrix $\boldsymbol{Q}_n$ such that (2.13) and (2.14) hold with $\boldsymbol{x}_t^{(n)} = \boldsymbol{Q}_n\boldsymbol{s}_t$. Then, this assumption and (3.8) together suggest a linear transformation, $\boldsymbol{F}_n$, of $\boldsymbol{P}_t$ that depends on G and $\boldsymbol{Q}_n$ and satisfies

$$\boldsymbol{F}_n\boldsymbol{P}_t = (\boldsymbol{x}_t^{(n)^\top}, \boldsymbol{v}_{t-1}^\top, \boldsymbol{z}_{t-1}^\top(1), \ldots, \boldsymbol{z}_{t-1}^\top(d))^\top,$$

in which each component has the same order of magnitude and the deterministic and random components are completely separated. Define $\boldsymbol{G}_t^{(n)} = \boldsymbol{F}_n\boldsymbol{P}_t$ and $\hat{\boldsymbol{R}}_n = n^{-1}\sum_{t=1}^n \boldsymbol{G}_t^{(n)}\boldsymbol{G}_t^{(n)^\top}$. Since $(\boldsymbol{v}_{t-1}^\top, \boldsymbol{z}_{t-1}^\top(1), \ldots, \boldsymbol{z}_{t-1}^\top(d))^\top$ is a subvector of $\boldsymbol{Y}_{t-1}$ defined in (2.9), it follows from Theorem 2.1 that for $0 < q < \gamma$,

$$\mathbf{E}\left(\lambda_{\min}^{-q}(\hat{\boldsymbol{R}}_n)\right) = O(1), \tag{3.9}$$

provided (2.10) holds with $m = d + 1$ and

$$\sup_{-\infty<t<\infty} \mathbf{E}|\varepsilon_t|^{2\gamma} + \sup_{-\infty<t<\infty} \max_{1\le l\le d} \mathbf{E}|\varepsilon_t(l)|^{2\gamma} < \infty, \tag{3.10}$$

for some $\gamma > 1$. Equation (3.9) plays an indispensable role in dealing with $\mathbf{E}(y_{n+1} - \hat{y}_{n+1})^2$ because it is not possible to rigorously analyze

$$n(\mathbf{E}(y_{n+1} - \hat{y}_{n+1})^2 - \sigma_{11}) = \mathbf{E}\big(\boldsymbol{G}_{n+1}^{(n)^\top}\hat{\boldsymbol{R}}_n^{-1} n^{-1/2}\sum_{t=1}^n \boldsymbol{G}_t^{(n)}\varepsilon_t\big)^2, \tag{3.11}$$

without recourse to the moment bounds associated with $\hat{\boldsymbol{R}}_n^{-1}$ or $\lambda_{\min}^{-1}(\hat{\boldsymbol{R}}_n)$. Recall $\sigma_{11} = \mathbf{E}(\varepsilon_t^2)$ defined after (2.7). The main result of this paper is stated in the next theorem.

Theorem 3.1 *Assume* (3.1), (3.3), (3.4), (2.10), *with* $m = d+1$, *and* (3.10), *with* $\gamma > 4$.

Also assume that there exists a $p \times p$ *non-random matrix* Q_n *such that* (2.13) *and* (2.14) *hold with* $\boldsymbol{x}_t^{(n)} = \boldsymbol{Q}_n \boldsymbol{s}_t$*. Then,*

$$
\begin{aligned}
n\left[\mathbf{E}(y_{n+1}-\hat{y}_{n+1})^2-\sigma_{11}\right]+o(1) &= \boldsymbol{x}_{n+1}^{(n)\top}(n^{-1}\sum_{t=1}^{n}\boldsymbol{x}_t^{(n)}\boldsymbol{x}_t^{(n)\top})^{-1}\boldsymbol{x}_{n+1}^{(n)}\sigma_{11} \\
&+\sigma_{11}\sum_{j=0}^{d}k_j.
\end{aligned} \tag{3.12}
$$

Ignoring the $o(1)$ term, the centered MSPE, $\mathbf{E}(y_{n+1}-\hat{y}_{n+1})^2-\sigma_{11}$, multiplied by the sample size, can be expressed as the sum of two terms. The second term, accounting for the variation due to estimating the ARX part of the model, is linearly proportional to the number of the estimated parameters. On the other hand, the first term (due to the error arising from estimating the deterministic trend) has asymptotic behavior varying appreciably, depending on the time trend's feature. The following examples provide more illustrations of Theorem 3.1.

Example 3.1 *Consider the polynomial trend* (3.6)*. Set* $\boldsymbol{Q}_n = \mathrm{Diag}(1, n^{-1}, \ldots, n^{-p+1})$*. Then,* $n^{-1}\sum_{t=1}^{n}\boldsymbol{x}_t^{(n)}\boldsymbol{x}_t^{(n)\top} \to \boldsymbol{H}_p = (1/(i+j-1))_{1\le i,j\le p}$*, the p-dimensional Hilbert matrix, and* $\boldsymbol{x}_{n+1}^{(n)} \to \mathbf{1}_p$*, the p-dimensional vector of ones. Since* $\boldsymbol{H}_p^{-1}$ *exists (see Choi (1983)), Theorem 3.1 implies*

$$
\lim_{n\to\infty} n\left[\mathbf{E}(y_{n+1}-\hat{y}_{n+1})^2-\sigma_{11}\right] = \sigma_{11}(\mathbf{1}_p^\top \boldsymbol{H}_p^{-1}\mathbf{1}_p + \sum_{j=0}^{d}k_j). \tag{3.13}
$$

Example 3.2 *For the periodic trend* (3.7)*, set* $\boldsymbol{Q}_n = I_{2h+1}$*. Then,*

$$\lim_{n\to\infty} n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_t^{(n)} \boldsymbol{x}_t^{(n)^\top} = \mathrm{Diag}(1, 1/2, \ldots, 1/2),$$

and $\boldsymbol{x}_{n+1}^{(n)} = (1, \sin\nu_1(n+1), \cos\nu_1(n+1) \ldots, \sin\nu_h(n+1), \cos\nu_h(n+1))^\top$*. Therefore,* $\boldsymbol{x}_{n+1}^{(n)^\top}(n^{-1}\sum_{t=1}^{n} \boldsymbol{x}_t^{(n)} \boldsymbol{x}_t^{(n)^\top})^{-1} \boldsymbol{x}_{n+1}^{(n)} \to 2h+1$*, and hence by Theorem 3.1,*

$$\lim_{n\to\infty} n \left[\mathbf{E}(y_{n+1} - \hat{y}_{n+1})^2 - \sigma_{11}\right] = \sigma_{11}(2h + 1 + \sum_{j=0}^{d} k_j). \tag{3.14}$$

Example 3.2 reveals that the impact of the periodic trend aligns with that of the ARX component. That is, it is linearly proportional to the number of parameters. On the other hand, Example 3.1 shows that this is not the case for the polynomial trend since $\mathbf{1}_p^\top \boldsymbol{H}_p^{-1} \mathbf{1}_p / p$ is not a constant. We will explore this issue further in the next section. When $d = 0$ (no exogenous variables in the model), (3.13) was given in Ing (2003) under the stringent condition that $\mathbf{E}|\varepsilon_t|^q < \infty$ for any $q > 0$. His derivation depends on Lemma B.1, which claims that (3.9) holds for any $q > 0$, provided the time trend satisfies (3.6). However, a rigorous proof of this result seems to be missing.

Before closing this section, we note that the relevance of our asymptotic results in finite samples has been investigated through a limited simulation study, which concentrates on several AR(2) models with polynomial or periodic trends. Our simulations show that the empirical MSPEs, obtained based on 10,000 replications, are quite close to their limiting values given on the right-hand sides of (3.13) and (3.14) even for $n = 100$. Further details are available upon request from the authors.

3.2 Statistical predictions and Cauchy matrices

To get a better understanding of the impact of the polynomial time trend on the corresponding MSPE, it is of interest to calculate the value of $\mathbf{1}_p^\top \boldsymbol{H}_p^{-1}\mathbf{1}_p$ in (3.13). In fact, $\boldsymbol{H}_p$ is a special case of the symmetric Cauchy matrix $\boldsymbol{C}_p = ((l_i + l_j - 1)^{-1})_{1\leq i,j\leq p}$, where $l_1 \dots l_p$ are distinct real numbers with $l_i + l_j \neq 1$ for all $1 \leq i, j \leq p$. In this section, we assume $\min_{1\leq i\leq p} l_i > 1/2$ to ensure that $\boldsymbol{C}_p$ is positive definite; see Fiedler (2010). Obviously, when $l_i = i, i = 1, \dots, p$, $\boldsymbol{C}_p = \boldsymbol{H}_p$. By making use of Lagrange's interpolation formula, Schechter (1959) showed that

$$\mathbf{1}_p^\top \boldsymbol{C}_p^{-1}\mathbf{1}_p = (\sum_{j=1}^{p} 2l_j) - p, \tag{3.15}$$

which can be applied to aerodynamics. Equation (3.15) leads immediately to

$$\mathbf{1}_p^\top \boldsymbol{H}_p^{-1}\mathbf{1}_p = p^2, \tag{3.16}$$

showing that estimating the polynomial trend yields a prediction error quadratically proportional to the number of parameters associated with the trend. This is in contrast to the prediction error contributed by estimating the ARX part, which is linearly proportional to the number of parameters. In view of the connection between $\mathbf{1}_p^\top \boldsymbol{H}_p^{-1}\mathbf{1}_p$ and statistical prediction, it is intriguing to explore if there exists a statistical proof of (3.16) or even (3.15). In the rest of this section, we show that the answer to this question is affirmative. Our proof of (3.15) ((3.16)) is reliant on a novel link between the MSPE and APE.

To begin with, let us focus on the following regression model,

$$y_t = \sum_{j=1}^{p} \beta_j t^{l_j - 1} + \varepsilon_t,\ t = 1, \dots, n \tag{3.17}$$

where $l_i > 1/2, i = 1, \dots, p$, and ε_t are independent standard normal random variables. The least squares predictor, $\hat{y}_{n+1}$, of y_{n+1} is given by $\boldsymbol{x}_{n+1}^\top \hat{\boldsymbol{\beta}}_n$, where

$$\hat{\boldsymbol{\beta}}_t = (\sum_{j=1}^{t} \boldsymbol{x}_j \boldsymbol{x}_j^\top)^{-1} \sum_{j=1}^{t} \boldsymbol{x}_j y_j,$$

with $\boldsymbol{x}_t = (t^{l_1-1}, \dots, t^{l_p-1})^\top$. Define $D_n = \text{Diag}(n^{l_1-1/2}, \dots, n^{l_p-1/2})$. Then, by the positive definiteness of $\boldsymbol{C}_p$, it can be shown that

$$\begin{aligned} &\lim_{n\to\infty} n\{\mathbf{E}(y_{n+1} - \hat{y}_{n+1})^2 - 1\} = \lim_{n\to\infty} n\{\mathbf{E}(y_{n+1} - \hat{y}_{n+1} - \varepsilon_{n+1})^2\} \\ &= \lim_{n\to\infty} n\boldsymbol{x}_{n+1}^\top D_n^{-1}(D_n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_t \boldsymbol{x}_t^\top D_n^{-1})^{-1} D_n^{-1} \boldsymbol{x}_{n+1} = \mathbf{1}_p^\top \boldsymbol{C}_p^{-1} \mathbf{1}_p, \end{aligned} \tag{3.18}$$

which establishes a connection between the left-hand side of (3.15) and the MSPE. The key idea is to further associate the MSPE in (3.18) with the APE, $\sum_{t=M+1}^{n}(y_t - \hat{y}_t - \varepsilon_t)^2$. More specifically, it follows from

$$\boldsymbol{x}_n^\top (\sum_{t=1}^{n} \boldsymbol{x}_n \boldsymbol{x}_n^\top)^{-1} \boldsymbol{x}_n \to 0, \quad \liminf_{n\to\infty} \lambda_{\min}(D_n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_n \boldsymbol{x}_n^\top D_n^{-1}) > 0,$$

Theorem 3 of Wei (1987), and the positive definiteness of $\boldsymbol{C}_p$ that

$$\lim_{n\to\infty} \frac{\sum_{t=M+1}^{n}(y_t - \hat{y}_t - \varepsilon_t)^2}{\log\det\sum_{t=1}^{n}\boldsymbol{x}_n\boldsymbol{x}_n^\top} = \lim_{n\to\infty}\frac{\sum_{t=M+1}^{n}(y_t - \hat{y}_t - \varepsilon_t)^2}{[(\sum_{j=1}^{p}2l_j) - p]\log n} = 1 \text{ a.s.}, \tag{3.19}$$

where M is the smallest integer t such that $\hat{\boldsymbol{\beta}}_t$ is uniquely defined. By Minkowski's inequality, it can be shown that $\{\sum_{t=M+1}^{n}(y_t - \hat{y}_t - \varepsilon_t)^2/\log n\}$ is uniformly integrable, which, together with (3.19), implies

$$\lim_{n\to\infty}\frac{1}{\log n}\sum_{t=M+1}^{n}\mathbf{E}(y_t - \hat{y}_t - \varepsilon_t)^2 = (\sum_{j=1}^{p}2l_j) - p. \tag{3.20}$$

Denote $\mathbf{E}(y_t - \hat{y}_t - \varepsilon_t)^2$ by ν_t. Then, (3.18) and (3.20) ensure

$$\lim_{n\to\infty} n\nu_n = \mathbf{1}_p^\top\boldsymbol{C}_p^{-1}\mathbf{1}_p \text{ and } \lim_{n\to\infty}\frac{1}{\log n}\sum_{t=M+1}^{n}\nu_t = (\sum_{j=1}^{p}2l_j) - p, \tag{3.21}$$

respectively. Moreover, it follows from the first identity of (3.21) that

$$\frac{1}{\log n}\sum_{t=M+1}^{n}\nu_t = \frac{1}{\log n}\left\{\sum_{t=M+1}^{n}t^{-1}(t\nu_t - \mathbf{1}_p^\top\boldsymbol{C}_p^{-1}\mathbf{1}_p) + \mathbf{1}_p^\top\boldsymbol{C}_p^{-1}\mathbf{1}_p\sum_{t=M+1}^{n}t^{-1}\right\}$$
$$= \mathbf{1}_p^\top\boldsymbol{C}_p^{-1}\mathbf{1}_p + o(1),$$

which, together with the second identity of (3.21), yields (3.15).

4 Conclusion

In this paper, we provide a rigorous analysis of the MSPE of the least squares predictor in ARX models with deterministic time trends satisfying some general conditions. Due to the difficulty in proving moment bounds for $\lambda_{\min}^{-q}(\hat{\boldsymbol{R}}_n), q \geq 1$, the asymptotic expression, (3.12), has not been reported elsewhere to the best of our knowledge. In the polynomial time trend, (3.12) inspires an intriguing proof of the formula for $\mathbf{1}_p^\top \boldsymbol{C}_p^{-1} \mathbf{1}_p$ from a prediction perspective. However, there are still several issues that require further investigation. For example, both the polynomial and periodic time trends, (3.6) and (3.7), are precluded by (3.4) if $1 - a_1 z - \cdots - a_{k_0} z^{k_0}$ has roots on the unit circle. This leads to the question on how to modify (3.13) and (3.14) in the presence of unit-roots. The techniques developed in Chan (1989) may be helpful to answer this question. Also, since the models imposed on the exogenous variables $z_t(l)$s are very general, the multistep prediction based on a finite number of lags of $z_t(l)$ may result in model misspecification. Therefore, an extension of (3.12) to the case of multistep prediction or model misspecification is another interesting topic for future research.

A Appendix

A.1 Proof of Theorem 2.1

To prove Theorem 2.1, we need some technical lemmas to characterize (2.14).

Lemma A.1 *Assume* (2.13). *Then,* (2.14) *holds if and only if there exists a subset,* G_n*,* *of* $\boldsymbol{X}_n = \{1, \ldots, n\}$*, with* $\liminf_{n\to\infty} \sharp(G_n)/n > 0$ *and* $\liminf_{n\to\infty} \min_{t\in G_n} \|\boldsymbol{x}_t^{(n)}\| > 0$ *such that*

$$\liminf_{n\to\infty} \lambda_{\min}(\sharp(G_n)^{-1} \sum_{t\in G_n} \boldsymbol{x}_t^{(n)} \boldsymbol{x}_t^{(n)^\top}) > 0. \tag{A.1}$$

PROOF. The proof of the "if" part of Lemma A.1 is easy and hence omitted. To prove the "only if" part, we note that (2.14) yields that for all large n,

$$\lambda_{\min}(n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_t^{(n)} \boldsymbol{x}_t^{(n)^\top}) > s, \tag{A.2}$$

where s is some positive constant, and hence

$$n^{-1} \sum_{t=1}^{n} x_{t,1}^2 > s, \tag{A.3}$$

where $x_{t,i}$ denotes the ith component of $\boldsymbol{x}_t^{(n)}$. Therefore, $G_n := \{t : 1 \le t \le n, \|\boldsymbol{x}_t^{(n)}\|^2 > s/2\}$ is non-empty for all large n. By (2.13) and (A.3), one has for all large n, $ns \le \sum_{t=1}^n x_{t,1}^2 \le \sharp(G_n)M_1^2 + ns/2$, yielding

$$\sharp(G_n) \ge sn/(2M_1^2). \tag{A.4}$$

Now, the desired conclusion follows from (A.4), (A.2), and $\lambda_{\min}(\sharp(G_n)^{-1} \sum_{t\in G_n} \boldsymbol{x}_t^{(n)} \boldsymbol{x}_t^{(n)^\top}) \ge \lambda_{\min}(n^{-1} \sum_{t\in G_n} \boldsymbol{x}_t^{(n)} \boldsymbol{x}_t^{(n)^\top}) \ge \lambda_{\min}(n^{-1} \sum_{t=1}^n \boldsymbol{x}_t^{(n)} \boldsymbol{x}_t^{(n)^\top}) - n^{-1} \sum_{t=1, t\notin G_n}^n \|\boldsymbol{x}_t^{(n)}\|^2$.

Lemma A.2 *Assume* (2.13)*. Then,* (2.14) *holds if and only if there exist disjoint subsets,*

$D_1, \ldots D_{q_n}$*, of* $\boldsymbol{X}_n$*, where* $\sharp(D_i) = p,\ i = 1, \ldots, q_n$*, and* $\liminf_{n\to\infty} q_n/n > 0$*, such that*

$$\liminf_{n\to\infty} \min_{1\le i\le q_n} \lambda_{\min}(\sum_{t\in D_i} \boldsymbol{x}_t^{(n)} \boldsymbol{x}_t^{(n)^\top}) > 0. \tag{A.5}$$

PROOF. The proof of the "if" part of Lemma A.2 is easy and hence omitted. To prove the "only if" part, by Lemma A.1 and (2.13), we can set $\|\boldsymbol{x}_t^{(n)}\| = 1$ for all t. We also assume without loss of generality that the s in (A.2) is less than 1. Define $D_0(n) = \emptyset$, and for $i \ge 1$, let $D_i(n)$ be any p-element subset of $\boldsymbol{X}_n - \bigcup_{l=0}^{i-1} D_l(n)$ satisfying

$$\lambda_{\min}\big(\sum_{t\in D_i(n)} \boldsymbol{x}_t^{(n)} \boldsymbol{x}_t^{(n)^\top}\big) > \frac{s^p}{2^p p^{p-1}}, \tag{A.6}$$

and $D_i(n) = \emptyset$ if no such subset is found. Also define $q_n = 0$ if $D_1(n) = \emptyset$, and $\max\{i \ge 1 : D_i(n) \ne \emptyset\}$ otherwise. For notational simplicity, in what follows we suppress the dependence of $\boldsymbol{x}_t^{(n)}$ and $D_i(n)$ on n, and write $\boldsymbol{x}_t$ and D_i instead of $\boldsymbol{x}_t^{(n)}$ and $D_i(n)$, respectively. Denote $\boldsymbol{X}_n - \bigcup_{l=0}^{q_n} D_l$ by $\boldsymbol{Z}_n = \{s_1, \ldots, s_{k_n}\}$, where $k_n = n - pq_n$. If $k_n < p$, then

$$q_n > n/p - 1. \tag{A.7}$$

For $k_n \ge p$, choose distinct elements, $c_1, \ldots, c_p$, in $\boldsymbol{Z}_n$ sequentially as follows. Let c_1 be any element of $\boldsymbol{Z}_n$. For $2 \le j \le p$, set $c_j = \arg\max_{s_i \in \boldsymbol{Z}_n} \|(I_p - M_{j-1})\boldsymbol{x}_{s_i}\|$, where M_{j-1} is the orthogonal projection matrix onto $C(\boldsymbol{x}_{c_1}, \ldots, \boldsymbol{x}_{c_{j-1}})$, the column space of $(\boldsymbol{x}_{c_1}, \ldots, \boldsymbol{x}_{c_{j-1}})$. Note that this sequential procedure of choosing c_i is similar in spirit to the orthogonal greedy algorithm in Ing and Lai (2011). Let M_0 be the $p \times p$ matrix of zeros. Then,

$$\|(I_p - M_{j-1})\boldsymbol{x}_{c_j}\| \text{ is non-increasing in } j, \tag{A.8}$$

and

$$\begin{aligned} &\prod_{j=1}^{p} \|(I_p - M_{j-1})\boldsymbol{x}_{c_j}\|^2 \leq \lambda_{\min}(\sum_{j=1}^{p} \boldsymbol{x}_{c_j}\boldsymbol{x}_{c_j}^\top)\lambda_{\max}^{p-1}(\sum_{j=1}^{p} \boldsymbol{x}_{c_j}\boldsymbol{x}_{c_j}^\top) \\ &\leq \lambda_{\min}(\sum_{j=1}^{p} \boldsymbol{x}_{c_j}\boldsymbol{x}_{c_j}^\top)p^{p-1} \leq (s/2)^p, \end{aligned} \tag{A.9}$$

where $\lambda_{\max}(A)$ denotes the maximum eigenvalue of A and the last inequality is ensured by

$$\max_{D \subset \boldsymbol{Z}_n, \sharp(D)=p} \lambda_{\min}\big(\sum_{t \in D} \boldsymbol{x}_t\boldsymbol{x}_t^\top\big) \leq \frac{s^p}{2^p p^{p-1}}.$$

Equations (A.8) and (A.9) imply there exists $1 < j^* \leq p$ such that for all $j^* \leq j \leq p$ and all large n, $\|(I_p - M_{j-1})\boldsymbol{x}_{c_j}\|^2 \leq s/2$, and hence for all $1 \leq i \leq k_n$ and all large n,

$$\|(I_p - M_{j^*-1})\boldsymbol{x}_{s_i}\|^2 \leq s/2. \tag{A.10}$$

Let $\boldsymbol{v}$ be any unit vector perpendicular to $C(M_{j^*-1})$. Then, (A.10) yields

$$\boldsymbol{v}^\top(n^{-1}\sum_{i=1}^{k_n}\boldsymbol{x}_{s_i}\boldsymbol{x}_{s_i}^\top)\boldsymbol{v} \le n^{-1}\sum_{i=1}^{k_n}[\boldsymbol{x}_{s_i}^\top(I_p - M_{j^*-1})\boldsymbol{v}]^2 \le s/2,$$

and hence $\lambda_{\min}(n^{-1}\sum_{i=1}^{k_n}\boldsymbol{x}_{s_i}\boldsymbol{x}_{s_i}^\top) \le s/2$. This, together with (A.2) and

$$\lambda_{\min}(n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top) = \lambda_{\min}(n^{-1}\sum_{i=0}^{q_n}\sum_{l\in D_i}\boldsymbol{x}_l\boldsymbol{x}_l^\top + n^{-1}\sum_{j=1}^{k_n}\boldsymbol{x}_{s_j}\boldsymbol{x}_{s_j}^\top)$$
$$\le \frac{q_n p}{n} + \lambda_{\min}(n^{-1}\sum_{j=1}^{k_n}\boldsymbol{x}_{s_j}\boldsymbol{x}_{s_j}^\top),$$

gives

$$q_n \ge \frac{sn}{2p}, \tag{A.11}$$

for all large n. Consequently, the desired conclusion follows from (A.7) and (A.11).

With the help of Lemma A.2, we are now in a position to prove Theorem 2.1.

PROOF OF THEOREM 2.1. Consider

$$\boldsymbol{\omega}_t^* = \begin{pmatrix} I_p & \mathbf{0}' \\ \mathbf{0} & C_0^{-1} \end{pmatrix}\begin{pmatrix} \boldsymbol{x}_t^{(n)} \\ \boldsymbol{Y}_t \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{x}_t^{(n)} \\ \boldsymbol{Y}_t^* \end{pmatrix},$$

where $\boldsymbol{Y}_t^* = \boldsymbol{\varepsilon}_{t,0} + \sum_{j=1}^{\infty} C_j^*\boldsymbol{\varepsilon}_{t,j}$, $C_j^* = C_0C_j$, and the dependence of $\boldsymbol{\omega}_t^*$ on n is suppressed in this proof. It follows from (2.2) that

$$\sum_{j=1}^{\infty} \|C_j^*\|_F^2 < \infty. \tag{A.12}$$

In the rest of the proof, we shall show that

$$\mathbf{E}\left(\lambda_{\min}^{-q}\big(n^{-1} \sum_{t=1}^{n} \boldsymbol{\omega}_t^* \boldsymbol{\omega}_t^{*^\top}\big) \right) = O(1), \tag{A.13}$$

which leads immediately to the desired result (2.12). By Lemma A.2, one has for all large n,

$$\min_{1 \le i \le q_n} \lambda_{\min} \left(\sum_{t \in D_i} \boldsymbol{x}_t^{(n)} \boldsymbol{x}_t^{(n)^\top} \right) > \rho_1, \tag{A.14}$$

where ρ_1 is some positive constant, D_is are disjoint subsets of $\boldsymbol{X}_n$ with $\sharp(D_i) = p$, for all $1 \le i \le q_n$, and q_n satisfies $\liminf_{n\to\infty} q_n/n > 0$. Let d_i denote the largest integer in D_i and $\{d_{(i)}\}$ be the decreasing rearrangement of $\{d_i\}$. Set $c_1 = d_{(1)}$ and for $i \ge 2$, define $c_i = \max\{d_{(l)}, 1 \le l \le q_n : c_{i-1} - d_{(l)} \ge k\}$, and 0 if no such $d_{(l)}$ exists. Let $s_n = \max\{i \ge 1 : c_i \ge 1\}$. Then, it is easy to see that $\liminf_{n\to\infty} s_n/n > 0$. Let $D_i^{'}$ denote the set $D_j, 1 \le j \le q_n$, containing c_i, L be an integer satisfying

$$L > \frac{2 + (q^{-1} + \gamma^{*^{-1}})\iota}{\alpha(q^{-1} - \gamma^{*^{-1}})}, \tag{A.15}$$

and $g_n = \lfloor s_n/L \rfloor$, where $q < \gamma^* < \gamma$, $\iota = p + km$, α is defined in (2.10), and $\lfloor a \rfloor$ is the largest integer $\le a$. Then, by the convexity of $x^{-q}, x > 0$, and $\liminf_{n\to\infty} g_n/n > 0$, we obtain for all large n,

$$\begin{aligned} &\text{the left-hand side of (A.13)} \leq \mathbf{E}\big(\lambda_{\min}^{-q}(n^{-1}\sum_{i=1}^{s_n}\sum_{t\in D_i'} \boldsymbol{B}_t)\big) \\ &\leq Cg_n^{-1}\sum_{j=0}^{g_n-1}\mathbf{E}\big(\lambda_{\min}^{-q}(\sum_{i=1}^{L}\sum_{t\in D_{i+jL}'} \boldsymbol{B}_t)\big), \end{aligned} \tag{A.16}$$

where $\boldsymbol{B}_t = \boldsymbol{\omega}_t^* \boldsymbol{\omega}_t^{*\top}$ and C here and hereafter represents a generic positive constant which is independent of n, but may vary from place to place. In view of (A.16), (A.13) follows if we can show that for all large n,

$$\mathbf{E}\big(\lambda_{\min}^{-q}(\sum_{i=1}^{L}\sum_{t\in D_{i+jL}'} \boldsymbol{B}_t)\big) \leq C,\ j=0,\ldots,g_n-1. \tag{A.17}$$

In the following, we only prove (A.17) for the case of $j=0$ because the proofs of other cases can be obtained similarly.

Let $k^* = \max\{k_1^*, k_2^*\}$, where k_1^* and k_2^* are positive constants to be specified later. Then, the left-hand sides of (A.17) (with $j=0$) is bounded by

$$\begin{aligned} &k^* + \int_{k^*}^{\infty} P(G(\mu)\bigcap V(\mu))d\mu + \int_{k^*}^{\infty} P\left(V^c(\mu)\right)d\mu \\ &:= k^* + \int_{k^*}^{\infty} \mathrm{I}(\mu)d\mu + \int_{k^*}^{\infty} \mathrm{II}(\mu)d\mu, \end{aligned} \tag{A.18}$$

where

$$G(\mu) = \left\{ \inf_{\substack{\|\boldsymbol{\nu}\|=1 \\ \boldsymbol{\nu}\in R^{\iota}}} \sum_{i=1}^{L}\sum_{t\in D_i'} (\boldsymbol{\nu}'\boldsymbol{\omega}_t^*)^2 < \mu^{-1/q} \right\},$$

and

$$V(\mu) = \left\{ \max_{t \in \bigcup_{i=1}^{L} D_i'} \|\boldsymbol{\omega}_t^*\|^2 \leq s\mu^{1/\gamma^*} \right\},$$

in which s is small enough such that

$$2M_1 p\sqrt{s} + ps \leq \rho_1/4, \tag{A.19}$$

where M_1 is defined in (2.13). By (2.15), (A.12), and Lemma 2 of Wei (1987), it can be shown that

$$\int_{k^*}^{\infty} \mathrm{II}(\mu) d\mu \leq C \int_{k^*}^{\infty} \mu^{-\gamma/\gamma^*} d\mu = O(1). \tag{A.20}$$

To deal with the first integration in (A.18), note that

$$\begin{aligned} G(\mu) \bigcap V(\mu) \subset \bigcap_{i=1}^{L} \{ \inf_{\substack{\|\boldsymbol{\nu}\|=1 \\ \boldsymbol{\nu} \in R^{\iota}}} \sum_{t \in D_i'} (\boldsymbol{\nu}' \boldsymbol{\omega}_t^*)^2 < \mu^{-1/q}, \sum_{t \in D_i'} \|\boldsymbol{\omega}_t^*\|^2 \leq ps\mu^{1/\gamma^*} \} \bigcap V(\mu) \\ := Q(\mu) \bigcap V(\mu). \end{aligned}$$

By an argument similar to that used in Ing and Wei (2003, page 137), it can be shown that there exist a positive integer $m^* \leq C_1 \mu^{\iota(q^{-1}+\gamma^{*^{-1}})/2}$ and unit vectors, $\boldsymbol{l}_1, \ldots, \boldsymbol{l}_m^*$, in R^{ι} such that

$$Q(\mu) \subset \bigcup_{j=1}^{m^*} \{ \|\boldsymbol{l}_j\|_{\sum_{t \in D_i'} \boldsymbol{B}_t} \leq (2\sqrt{ps\iota} + 1)\mu^{-1/2q},\ i = 1, \ldots, L \} := \bigcup_{j=1}^{m^*} \Pi_{j,L}(\mu),$$

where C_1 is some positive constant independent of n and μ and $\|\boldsymbol{x}\|_A^2 = \boldsymbol{x}^\top A \boldsymbol{x}$ for nonnegative definite matrix A. Since $|\boldsymbol{l}_j^\top \boldsymbol{\omega}^*_{c_i}| \le \|\boldsymbol{l}_j\|_{\sum_{t\in D_i'} \boldsymbol{B}_t}$,

$$
\begin{aligned}
\mathrm{I}(\mu) \le & \sum_{\substack{j=1 \\ \|\boldsymbol{l}_{j,2}\| \ge \mu^{-1/(2\gamma^*)}}}^{m^*} P\left(|\boldsymbol{l}_j^\top \boldsymbol{\omega}^*_{c_i}| \le (2\sqrt{ps\iota}+1)\mu^{-1/2q},\ i=1,\ldots,L\right) \\
+ & \sum_{\substack{j=1 \\ \|\boldsymbol{l}_{j,2}\| < \mu^{-1/(2\gamma^*)}}}^{m^*} P\left(V(\mu), \Pi_{j,L}(\mu)\right) := \sum_{\substack{j=1 \\ \|\boldsymbol{l}_{j,2}\| \ge \mu^{-1/(2\gamma^*)}}}^{m^*} \mathrm{III}_j(\mu) \\
+ & \sum_{\substack{j=1 \\ \|\boldsymbol{l}_{j,2}\| < \mu^{-1/(2\gamma^*)}}}^{m^*} \mathrm{IV}_j(\mu),
\end{aligned}
\tag{A.21}
$$

where $\boldsymbol{l}_{j,2}$ is the vector consisting of $\boldsymbol{l}_j$'s last km elements. Denote $\boldsymbol{l}_{j,2}$ by $(\boldsymbol{l}_{j,2}^\top(1),\ldots,\boldsymbol{l}_{j,2}^\top(k))^\top$, where each of $\boldsymbol{l}_{j,2}(i)$ is m-dimensional. Then, for $\|\boldsymbol{l}_{j,2}\| \ge \mu^{-1/(2\gamma^*)}$ and $\mu \ge k_1^* = \{2\sqrt{k}(2\sqrt{ps\iota}+1)/\delta\}^{2/(q^{-1}-\gamma^{*-1})}$, it holds that

$$
\begin{aligned}
\mathrm{III}_j(\mu) &= \mathbf{E}\left\{\prod_{i=2}^{L} I_{\{|\boldsymbol{l}_j^\top \boldsymbol{\omega}^*_{c_i}|\le(2\sqrt{ps\iota}+1)\mu^{-1/2q}\}} P\left(\left.\boldsymbol{A}_1(\mu)\right|\boldsymbol{\delta}_{c_1-j}, j\ge k\right)\right\} \\
&\le M\{2\sqrt{k}(2\sqrt{ps\iota}+1)\mu^{-(q^{-1}-\gamma^{*-1})/2}\}^{\alpha}\mathbf{E}\left(\prod_{i=2}^{L} I_{\{|\boldsymbol{l}_j^\top \boldsymbol{\omega}^*_{c_i}|\le(2\sqrt{ps\iota}+1)\mu^{-1/2q}\}}\right)
\end{aligned}
$$

where $\boldsymbol{A}_1(\mu) = \{-(2\sqrt{ps\iota}+1)\mu^{-1/2q} - r_1^* \le \sum_{h=1}^{k} \boldsymbol{l}_{j,2}^\top(h)\boldsymbol{\delta}_{c_1+1-h} \le (2\sqrt{ps\iota}+1)\mu^{-1/2q} - r_1^*\}$, with $r_1^* = \boldsymbol{l}_{j,1}^\top \boldsymbol{x}_{c_1}^{(n)} + \boldsymbol{l}_{j,2}^\top \sum_{j=1}^{\infty} C_j^* \boldsymbol{\varepsilon}_{c_1,j}$ and $(\boldsymbol{l}_{j,1}^\top, \boldsymbol{l}_{j,2}^\top)^\top = \boldsymbol{l}_j$, and the first inequality follows from $\|\boldsymbol{l}_{j,2}(h)\| \ge k^{-1/2}\mu^{-1/(2\gamma^*)}$ for some $1 \le h \le k$, (2.10), and the independence among

$\{\boldsymbol{\delta}_t\}$. Repeat the same argument $L-1$ times, one gets,

$$\mathrm{III}_j(\mu) \leq M^L\{2\sqrt{k}(2\sqrt{ps\iota}+1)\}^{vL}\mu^{-(q^{-1}-{\gamma^*}^{-1})\alpha L/2},$$

provided $\|\boldsymbol{l}_{j,2}\| \geq \mu^{-1/(2\gamma^*)}$ and $\mu \geq k_1^*$. As a result, by (A.15),

$$\int_{k^*}^{\infty} \sum_{\substack{j=1 \\ \|\boldsymbol{l}_{j,2}\| \geq \mu^{-1/(2\gamma^*)}}}^{m^*} \mathrm{III}_j(\mu) d\mu \leq C \int_{k^*}^{\infty} \mu^{-[(q^{-1}-{\gamma^*}^{-1})\alpha L/2-\iota(q^{-1}+{\gamma^*}^{-1})/2]} = O(1). \tag{A.22}$$

For $\|\boldsymbol{l}_{j,2}\| < \mu^{-1/(2\gamma^*)}$ and $\mu \geq k_2^* = \max\{2^{\gamma^*}, \{5(2\sqrt{ps\iota}+1)^2/\rho_1\}^q\}$, (A.14) and (A.19) ensure that on the set $V(\mu)$,

$$\begin{aligned}
&\min_{1\leq i\leq L} \|\boldsymbol{l}_j\|_{\sum_{t\in D_i'}\boldsymbol{B}_t} \geq (\min_{1\leq i\leq L} \|\boldsymbol{l}_{j,1}\|^2_{\sum_{t\in D_i'}\boldsymbol{x}_t^{(n)}\boldsymbol{x}_t^{(n)\top}} - 2M_1p\sqrt{s} - ps)^{1/2} \\
&\geq (\rho_1/2 - 2M_1p\sqrt{s} - ps)^{1/2} \geq (\rho_1/4)^{1/2} > (\rho_1/5)^{1/2} \geq (2\sqrt{ps\iota}+1)\mu^{-1/2q},
\end{aligned}$$

for all large n. Hence, for all large n,

$$\int_{k^*}^{\infty} \sum_{\substack{j=1 \\ \|\boldsymbol{l}_{j,2}\| < \mu^{-1/(2\gamma^*)}}}^{m^*} \mathrm{IV}_j(\mu) d\mu = 0. \tag{A.23}$$

Consequently, (A.17) (with $j=0$) follows from (A.18), (A.20), (A.21), (A.22), and (A.23). Thus, the proof is complete.

A.2 Proof of Theorem 3.1

PROOF OF THEOREM 3.1. We can assume without loss of generality that $\hat{\boldsymbol{R}}_n^{-1}$ exists because (3.9) implies $P(\hat{\boldsymbol{R}}_n^{-1}$ exists$) = 1$ for all large n. Denote $(\boldsymbol{v}_t^\top(k_0), \boldsymbol{z}_t^\top(k_1), \ldots, \boldsymbol{z}_t^\top(k_d))^\top$ by $\boldsymbol{Q}_t = (Q_t(1), \ldots, Q_t(\sum_{l=0}^d k_l))^\top$ and $\mathbf{E}(\boldsymbol{Q}_t\boldsymbol{Q}_t^\top)$ by $\boldsymbol{F} = (F_{i,j})_{1\leq i,j\leq \sum_{l=0}^d k_l}$. Then, it follows from (2.6), (2.7), and the first moment bound theorem of Findley and Wei (1993) that for any $1 \leq i, j \leq \sum_{l=0}^d k_l$,

$$\mathbf{E}\big(n^{-1}\sum_{t=1}^n Q_t(i)Q_t(j) - F_{i,j}\big)^2 \leq C\sqrt{\sum_{l=0}^{n-1}\gamma_l^2(i)/n}\sqrt{\sum_{l=0}^{n-1}\gamma_l^2(j)/n} = o(1), \tag{A.24}$$

where $\gamma_l(i)$ is the autocovariance function of $\{Q_t(i)\}$ at lag l. In addition, it is easy to see that for any $1 \leq i \leq \sum_{l=0}^d k_l$,

$$\mathbf{E}\|n^{-1}\sum_{t=1}^n \boldsymbol{x}_t^{(n)} Q_{t-1}(i)\|^2 \to \mathbf{0},$$

which, together with (A.24), yields

$$\hat{\boldsymbol{R}}_n - \begin{pmatrix} n^{-1}\sum_{t=1}^{n-1}\boldsymbol{x}_t^{(n)}\boldsymbol{x}_t^{(n)^\top} & \mathbf{0}_{p\times\sum_{l=0}^d k_l} \\ \mathbf{0}_{(\sum_{l=0}^d k_l)\times p} & \boldsymbol{F} \end{pmatrix} = o_p(1). \tag{A.25}$$

In view of (A.25), the desired conclusion follows if

$$\boldsymbol{T}_n := \big(\boldsymbol{G}_{n+1}^{(n)^\top}\hat{\boldsymbol{R}}_n^{-1} n^{-1/2}\sum_{t=1}^n \boldsymbol{G}_t^{(n)}\varepsilon_t\big)^2 \text{ is uniformly integrable}, \tag{A.26}$$

and

$$\begin{aligned}&\mathbf{E}\left(\boldsymbol{x}_t^{(n)^\top}(n^{-1}\sum_{t=1}^{n-1}\boldsymbol{x}_t^{(n)}\boldsymbol{x}_t^{(n)^\top})^{-1}n^{-1/2}\sum_{t=1}^{n}\boldsymbol{x}_t^{(n)}\varepsilon_t+\boldsymbol{Q}_n^\top\boldsymbol{F}^{-1}n^{-1/2}\sum_{t=1}^{n}\boldsymbol{Q}_{t-1}\varepsilon_t\right)^2\\&=\boldsymbol{x}_{n+1}^{(n)^\top}(n^{-1}\sum_{t=1}^{n}\boldsymbol{x}_t^{(n)}\boldsymbol{x}_t^{(n)^\top})^{-1}\boldsymbol{x}_{n+1}^{(n)}\sigma_{11}+\sigma_{11}\sum_{j=0}^{d}k_j+o(1).\end{aligned}\tag{A.27}$$

Since $\gamma > 4$, there exists $\theta > 0$ small enough such that $4 < 2\gamma(1+\theta)/(\gamma-2(1+\theta)) < \gamma$. Let $2\gamma(1+\theta)/(\gamma-2(1+\theta)) \leq q^* < \gamma$. Then, by Theorem 2.1 and (3.10),

$$\mathbf{E}\{\lambda_{\min}^{-q^*}(\hat{\boldsymbol{R}}_n)\} = O(1). \tag{A.28}$$

By (A.28), Lemma 2 of Wei (1987), $4q^*(1+\theta)/(q^*-2(1+\theta)) \leq 2\gamma$, and Hölder's inequality,

$$\begin{aligned}\mathbf{E}(\boldsymbol{T}_n^{1+\theta}) \leq &\left(\mathbf{E}\|\boldsymbol{G}_{n+1}^{(n)}\|^{\frac{4q^*(1+\theta)}{q^*-2(1+\theta)}}\right)^{\frac{q^*-2(1+\theta)}{2q^*}}\left(\mathbf{E}\|n^{-1/2}\sum_{t=1}^{n}\boldsymbol{G}_t^{(n)}\varepsilon_t\|^{\frac{4q^*(1+\theta)}{q^*-2(1+\theta)}}\right)^{\frac{q^*-2(1+\theta)}{2q^*}}\\ &\times\left(\mathbf{E}\lambda_{\min}^{-q^*}(\hat{\boldsymbol{R}}_n)\right)^{2(1+\theta)/q^*} = O(1),\end{aligned}$$

leading to (A.26).

To prove (A.27), define $\tilde{\boldsymbol{Q}}_n = \mathbf{E}(\boldsymbol{Q}_n|\mathcal{F}_{n-g_n})$, where $g_n \to \infty$, $g_n = o(n)$, and $\mathcal{F}_t$ is the

σ-field generated by $(\boldsymbol{\delta}_t, \boldsymbol{\delta}_{t-1}, \ldots)$. Then,

$$\begin{aligned}\mathbf{E}\left(\boldsymbol{Q}_n^\top \boldsymbol{F}^{-1} n^{-1/2} \sum_{t=1}^{n} \boldsymbol{Q}_{t-1}\varepsilon_t\right)^2 &= \mathbf{E}\left(\boldsymbol{Q}_n^{*^\top} \boldsymbol{F}^{-1} n^{-1/2} \sum_{t=1}^{n-g_n} \boldsymbol{Q}_{t-1}\varepsilon_t\right)^2 \\ + \mathbf{E}\left(\tilde{\boldsymbol{Q}}_n^\top \boldsymbol{F}^{-1} n^{-1/2} \sum_{t=n-g_n+1}^{n} \boldsymbol{Q}_{t-1}\varepsilon_t\right)^2 &+ \mathbf{E}\left(\boldsymbol{Q}_n^{*^\top} \boldsymbol{F}^{-1} n^{-1/2} \sum_{t=n-g_n+1}^{n} \boldsymbol{Q}_{t-1}\varepsilon_t\right)^2 \\ + \mathbf{E}\left(\tilde{\boldsymbol{Q}}_n^\top \boldsymbol{F}^{-1} n^{-1/2} \sum_{t=1}^{n-g_n} \boldsymbol{Q}_{t-1}\varepsilon_t\right)^2 &:= \text{(I)} + \text{(II)} + \text{(III)} + \text{(IV)},\end{aligned} \tag{A.29}$$

where $\boldsymbol{Q}_n^* = \boldsymbol{Q}_n - \tilde{\boldsymbol{Q}}_n$. By Lemma 2 of Wei (1987) and the Cauchy-Schwarz inequality,

$$\text{(II)} + \text{(III)} + \text{(IV)} = o(1). \tag{A.30}$$

Straightforward calculations give

$$\lim_{n\to\infty} \text{(I)} = \sigma_{11} \sum_{j=0}^{d} k_j, \tag{A.31}$$

and

$$\begin{aligned}&\mathbf{E}\left(\boldsymbol{x}_t^{(n)^\top} (n^{-1} \sum_{t=1}^{n-1} \boldsymbol{x}_t^{(n)} \boldsymbol{x}_t^{(n)^\top})^{-1} n^{-1/2} \sum_{t=1}^{n} \boldsymbol{x}_t^{(n)} \varepsilon_t\right)^2 \\ &= \boldsymbol{x}_{n+1}^{(n)^\top} (n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_t^{(n)} \boldsymbol{x}_t^{(n)^\top})^{-1} \boldsymbol{x}_{n+1}^{(n)} \sigma_{11}\end{aligned} \tag{A.32}$$

An argument similar to that used to prove (A.30) yields

$$\mathbf{E}\left(n^{-1}\{\boldsymbol{x}_t^{(n)\top}(n^{-1}\sum_{t=1}^{n-1}\boldsymbol{x}_t^{(n)}\boldsymbol{x}_t^{(n)\top})^{-1}\sum_{t=1}^{n}\boldsymbol{x}_t^{(n)}\varepsilon_t\}(\boldsymbol{Q}_n^{\top}\boldsymbol{F}^{-1}\sum_{t=1}^{n}\boldsymbol{Q}_{t-1}\varepsilon_t)\right) = o(1). \qquad \text{(A.33)}$$

Consequently, (A.27) follows from (A.29)–(A.33). Thus the proof is complete.

Acknowledgement

We would like to thank the Associate Editor and two anonymous reviewers for their comments and suggestions that helped improve the readability of this paper. Ing's and Yu's research was supported in part by grants 109-2118-M-007-007-MY3 and 109-2118-M-390-002 from the Ministry of Science and Technology (MOST), Taiwan.

References

Bhansali, R. J. (1996). Asymptotically efficient autoregressive model selection for multistep prediction. *Annals of the Institute of Statistical Mathematics* **48** 577–602.

Chan, N.H. (1989). Asymptotic inference for unstable autoregressive time series with drifts. *Journal of Statistical Planning and Inference* **23** 301–312.

Chan, N.H. and Ing, C.-K. (2011). Uniform moment bounds of Fisher's information with applications to time series. *Annals of Statistics* **39** 1526–1550.

Chan, N.H., Huang, S.-F. and Ing, C.-K. (2013). Moment bound and mean squared prediction errors of long-memory time series. *Annals of Statistics* **41** 1268–1298.

Chen, H.-F. and Guo, Lei (1986). Convergence rate of least-squares identification and adaptive control for stochastic systems. *International Journal of Control* **44**, 1459–1476.

Chen, K., Hu, I., and Ying, Z. (1999). Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *Annals of Statistics* **27** 1155–1163.

Choi, M.-D. (1983) Tricks or treats with the Hilbert matrix. *American Mathematical Monthly* **90**, 301–312.

Fiedler, M. (2010). Notes on Hilbert and Cauchy matrices. *Linear Algebra and its Applications* **432** 351–356.

Findley, D.F. and Wei, C.-Z. (2002). AIC, overfitting principles, and the boundedness of moments of inverse matrices for vector autoregressions and related models. *Journal of Multivariate Analysis* **83** 415–450.

Findley, D.F. And Wei, C.-Z. (1993). Moment bounds for deriving time series CLTs and model selection procedures. *Statistica Sinica* **3** 453–480.

Fuller, W.A. and Hasza, D.P. (1981). Properties of predictors for autoregressive time series. *Journal of the American Statistical Association* **76** 155–161.

Gerencsér, L. (1992). AR(∞) estimation and nonparametric stochastic complexity. *IEEE Transactions on Information Theory* **38** 1768–1778.

Gerencsér, L., Hjalmarsson, H., and Mårtensson, J. (2009). Identification of ARX systems with non-stationary inputs–asymptotic analysis with application to adaptive input design. *Automatica* **45** 623–633.

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

Hannan, E.J., Mcdougall, A.J. and Poskit, D.S. (1989). Recursive estimation of autoregressions. *Journal of the Royal Statistical Society: Series B (Methodological)* **51** 217–233.

Hemerly, E.M. and Davis, M.H.A. (1989). Strong consistency of the PLS criterion for order determination of autoregressive processes. *Annals of Statistics* **17** 941–946.

Hsu, H.-L., Ing, C.-K., and Tong, H. (2019). On model selection from a finite family of possibly misspecified time series models. *Annals of Statistics* **47** 1061–1087.

Ing, C.-K. (2003). Multistep prediction in autoregressive processes. *Econometric Theory* **19** 254–279.

Ing, C.-K. (2004). Selecting optimal multistep predictors for autoregressive processes of unknown order. *Annals of Statistics* **32** 693–722.

Ing, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Annals of Statistics* **35** 1238–1277.

Ing, C.-K and Lai, T.L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica* **21** 1473–1513.

Ing, C.-K., Lin, J.-L, and Yu, S.-H. (2009). Toward optimal multistep forecasts in nonstationary autoregressions. *Bernoulli* **15** 402–437.

Ing, C.-K., Sin, C. Y. and Yu, S.-H. (2012). Model selection for integrated autoregressive processes of infinite order. *Journal of Multivariate Analysis* **106** 57–71.

Ing, C.-K. and Wei, C.-Z. (2003). On same-realization prediction in an infinite-order autoregressive process. *Journal of Multivariate Analysis* **85** 130-155.

Ing, C.-K. and Wei, C.-Z. (2005). Order selection for same-realization predictions in autoregressive processes. *Annals of Statistics* **33** 2423–2474.

Ing, C.-K. and Yang, C.-Y. (2014). Predictor selection for positive autoregressive processes. *Journal of the American Statistical Association* **109** 243–253.

Kunitomo, N. and Yamamoto, T. (1985). Properties of predictors in misspecified autoregressive time series. *Journal of the American Statistical Association* **80** 941–950.

Lai, T.L. (1994). Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *Annals of Statistics* **22** 1917–1930.

Lai, T.L. and Lee, C.P. (1997). Information and prediction criteria for model selection in stochastic regression and ARMA models. *Statistica Sinica* **7** 285–309.

Lai, T.L. and Wei, C.-Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics* **10** 154–166.

Lai, T.L. and Wei, C.-Z. (1986). Extended least squares and their applications to adaptive control and prediction in linear systems. *IEEE Transactions on Automatic Control* **AC-31** 898–906.

Lee, S. and Karagrigoriou, A. (2001). An asymptotically optimal selection of the order of a linear process. *Sankhyā: The Indian Journal of Statistics, Series A* **63** 93–106.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics* **14** 1080–1100.

Schechter, S. (1959). On the Inversion of Certain Matrices. *Mathematical Tables and Other Aids to Computation* **13** 73–77.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* **8** 147–164.

Speed, T.P. and Yu, B. (1993). Model selection and prediction: Normal regression. *Annals of the Institute of Statistical Mathematics* **45** 35–54.

Stock, J.H. (1994). Unit roots, structural breaks and trends, in R.F. Engle and D.L. McFadden eds., *Handbook of Econometrics, Volume IV*. North Holland, New York.

Wax, M. (1988). Order selection for AR models by predictive least squares. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **36** 581–588.

Wei, C.-Z. (1987). Adaptive prediction by least squares predictors in stochastic regression models with application to time series. *Annals of Statistics* **15** 1667–1682.

Wei, C.-Z. (1992). On predictive least squares principles. *Annals of Statistics* **20** 1–42.

West, K.D. (1996). Asymptotic inference about predictive ability. *Econometrica* **64** 1067–1084.