Consistency results for stationary autoregressive processes with constrained coefficients

Alessio Sancetta[*]

September 10, 2018

Abstract

We consider stationary autoregressive processes with coefficients restricted to an ellipsoid. These are included in the family of autoregressive processes with absolutely summable coefficients. We provide consistency results under different norms for the estimation of such processes using constrained and penalized estimators. As an application we show a weak form of universal consistency. Simulations show that directly including the constraint in the estimation can lead to more robust results.

Key Words: consistency, empirical process, ridge regression, reproducing kernel Hilbert space, universal consistency.

1 Introduction

It is common to impose constraints on the decay rate of the coefficients of the autoregression, in order to derive results amenable to estimation for the purpose of prediction. At minimum, these constraints usually require the coefficients of the autoregression to be absolutely summable. Then, a natural approach is to consider sieve estimation. Sieve estimation of infinite autoregressive (AR) models has been considered by various authors. For universal consistency, Schäfer (2002) derived perhaps the strongest result possible. Györfi and Sancetta (2015) review some of these results. For convergence in

[*] Acknowledgements: I am very grateful to the detailed comments of the referees and the editor that have lead to substantial improvements in content and presentation. E-mail: <asancetta@gmail.com>, URL: <http://sites.google.com/site/wwwsancetta/>. Address for correspondence: Department of Economics, Royal Holloway University of London, Egham TW20 0EX, UK

probability, various authors have considered infinite AR models and its applications, e.g. Bühlmann (1997), and Kreiss et al. (2011). Additional references can be found in the cited papers.

Here, we constrain the autoregression coefficients to lie inside an infinite dimensional ellipsoid such that the coefficients associated to higher order lags decay fast. The conditions essentially require the coefficients of the autoregression to be absolutely summable. We shall see that the vector of these coefficients can be seen as an element in a Reproducing Kernel Hilbert Space (RKHS) when ℓ_2 (the space of square summable sequences) is equipped with a suitable inner product. This allows us to exploit all the existing machinery for estimation in RKHS and build on it (Steinwart and Chirstmann, 2008, for a comprehensive review). The main ingredient is penalized least square estimation. We also consider the constrained least square problem. Penalized and constrained estimation are dual problems for specific values of the penalty coefficient. Our result establishes the relation between the two problems and the consistency rates. In general, they can lead to different consistency results under different norms. One norm is the usual ℓ_2 norm, while the other is the norm of the RKHS. We show that consistency under the latter has important implications for prediction problems.

In general, unlike existing results, we are able to establish consistency when both the order of the autoregression and the sample size go to infinity with no constraint on their rate of divergence to infinity. Existing results use the machinery of method of sieve, hence they require the order of the autoregression to go to infinity in a controlled way. However, the ellipsoid is compact under the ℓ_2 norm (Kuelbs, 1976, Example 2). Hence, we can expect to derive asymptotic results with no constraint on the growth rate of the number of estimated coefficients.

The plan for the paper is as follows. Section 2 reviews the estimation method and presents the consistency results. A numerical example is provided in Section 3. Section 4 mentions extensions to other processes such as vector autoregressive processes. The proof of the consistency results is long and is given in Section 5.

2 Estimation Method

We restrict attention to the stationary infinite order autoregressive process

$$Y_t = \sum_{k=1}^{\infty} \varphi_k Y_{t-k} + \varepsilon_t \tag{1}$$

for some mean zero independent identically distributed (i.i.d.) sequence $(\varepsilon_t)_{t\in\mathbb{Z}}$ with finite fourth moment, and unknown coefficients φ_k's. Throughout, $\mathbb{Z}$ is the set of integers and $\mathbb{N}$ is the set of strictly positive integers. Under additional conditions the model satisfies the following.

Lemma 1 *If (1) is such that $\sum_{k=1}^{\infty}|\varphi_k| < \infty$, and if $1-\sum_{k=1}^{\infty}\varphi_k z^k = 0$ only for z outside the unit circle, $(Y_t)_{t\in\mathbb{Z}}$ is stationary and ergodic with absolutely summable autocovariance function and $\mathbb{E}Y_t^4 < \infty$.*

It is well known that for an AR process, $1-\sum_{k=1}^{\infty}\varphi_k z^k = 0$ only for z outside the unit circle if the autocovariance function is absolutely summable and the spectral density is strictly positive and continuous (Kreiss et al., 2011, Corollary 2.1).

There are processes (even Gaussian) that satisfy the conditions of Lemma 1, but fail to be beta mixing (Doukhan, 1995, Theorem 3, p.59). The beta mixing assumption is often conveniently used when proving convergence using methods from empirical process theory. Alas, it cannot be used here.

In a finite sample, (1) can only be approximated by the finite dimensional model

$$Y_t = \sum_{k=1}^{K} b_k Y_{t-k} + \varepsilon_t$$

While this is essentially a sieve we do not necessarily require K to be of smaller order than the sample size. Here, we restrict the coefficients in the ellipsoid to be defined as follows. Let $\lambda = (\lambda_k)_{k\in\mathbb{N}}$ be a sequence of positive constants diverging to infinity. Define the ellipsoids

$$\mathcal{E} = \left\{ b \in \mathbb{R}^{\infty} : \sum_{k=1}^{\infty} b_k^2\lambda_k^2 < \infty \right\}$$

$$\mathcal{E}(B) := \left\{ b \in \mathbb{R}^{\infty} : \sum_{k=1}^{\infty} b_k^2\lambda_k^2 \leq B^2 \right\}$$

$$\mathcal{E}_K(B) := \left\{ b \in \mathbb{R}^{\infty} : \sum_{k=1}^{\infty} b_k^2\lambda_k^2 \leq B^2,\ b_k = 0 \text{ for } k > K \right\}. \tag{2}$$

Also, define the following subspace of $\mathbb{R}^{\infty}$ which will be used in defining a penalized estimator,

$$\mathcal{E}_K := \left\{ b \in \mathbb{R}^{\infty} : b_k = 0 \text{ for } k > K \right\}.$$

The ellipsoids depend on the choice of λ, which is supposed to be fixed throughout. For notational convenience, we do not make explicit the dependence on λ in the notation. When dealing with finite sample size, estimation is carried out in (2). We shall use the symbol $\lesssim$ to mean that the left hand side (l.h.s.) is less than an absolute constant times the right hand side (r.h.s.). When the inequality holds in both directions (with possibly different constants) we use the symbol $\asymp$.

We restrict the coefficients to be in $\mathcal{E}$ and work with the model in (1). Summarizing, we use the following conditions.

Condition 1 *For the sequence $(Y_t)_{t\in\mathbb{Z}}$ in (1), the following are satisfied:*

1. *(Stationarity) $1-\sum_{k=1}^{\infty}\varphi_k z^k=0$ only for z outside the unit circle;*
2. *(Parameter Space Constraint)$\varphi\in\mathcal{E}$ and the sequence λ is such that $\lambda_k\asymp k^{\eta}$, for $k\in\mathbb{N}$, where $\eta>1/2$;*
3. *(Innovations) the innovations $(\varepsilon_t)_{t\in\mathbb{Z}}$ are independent identically distributed with finite fourth moment.*

We use $\lambda_k\asymp k^{\eta}$ as in some applications (e.g. in the presence of seasonal patterns) we may wish to penalize the coefficients of the autoregression differently. For asymptotic analysis we may just set $\lambda_k=k^{\eta}$. Throughout, when writing $\mathcal{E}$ and similar quantities, it is understood that λ is as in Condition 1. The space $\mathcal{E}$ and hence λ are given. The only free parameter is B. The following is stated for convenience.

Lemma 2 *If $b\in\mathcal{E}(B)$ then, $|b_k|\leq c_B k^{-(2\eta+1)/2}/\sqrt{\ln(1+k)}$, where c_B is a finite constant that depends on B.*

In consequence, Condition 1 implies absolute summability of the coefficients of the autoregression and Lemma 1 applies. Absolute summability would just require $\eta\geq 1/2$ in Condition 1 rather than $\eta>1/2$. Hence, the condition we use is a bit more restrictive.

2.1 Estimation and Consistency

The goal is to find an estimator for φ. We consider two approaches: constrained least square and penalized least square. By duality, the two can be made to be equivalent by suitable choice of the penalty parameter. However, in the constrained case, the penalty turns out to be sample dependent, while in penalized estimation this it not necessarily the case.

To avoid notational trivialities, suppose that the sample size is $N = n + K$. This will be assumed without further notice throughout the paper. In particular, our sample is $Y_{-(K-1)}, Y_{-(K-2)}, ..., Y_0, Y_1, ..., Y_n$. This also stresses the fact that n and K can go to infinity at different rates.

In the constrained problem, we estimate $b \in \mathcal{E}_K(B)$. The constrained estimator is defined as

$$b^{(n)} = \arg \inf_{b \in \mathcal{E}_K(B)} \frac{1}{n} \sum_{t=1}^{n} \left(Y_t - \sum_{k=1}^{\infty} b_k Y_{t-k} \right)^2 \tag{3}$$

Of course, in the above, $\sum_{k=1}^{\infty} b_k Y_{t-k} = \sum_{k=1}^{K} b_k Y_{t-k}$ if $b \in \mathcal{E}_K(B)$.

For the penalized problem 1 define

$$b^{(n,\tau)} := \arg \inf_{b \in \mathcal{E}_K} \frac{1}{n} \sum_{t=1}^{n} \left(Y_t - \sum_{k=1}^{\infty} b_k Y_{t-k} \right)^2 + \tau \sum_{k=1}^{\infty} \lambda_k^2 b_k^2, \tag{4}$$

where $(\lambda_k)_{k \in \mathbb{N}}$ is as in the definition of $\mathcal{E}$, and $\tau > 0$. By use of the Lagrangian, we can always rewrite (3) as (4) for suitable choice of τ. This means that there is a $\tau = \tau_{B,n}$ ($\tau = 0$ if the constraint it not binding) such that $b^{(n,\tau)} = b^{(n)}$.

Both problems can be reformulated using matrix notation. Let X be the $n \times K$ dimensional matrix with $(t,k)^{th}$ entry equal to Y_{t-k} and Y be the n-dimensional vector with t^{th} entry Y_t. Also, let Λ be the $K \times K$ diagonal matrix with k^{th} diagonal entry equal to λ_k. The estimator for either (3) or (4) is found by minimizing the penalized least square criterion with respect to (w.r.t.) $\tilde{b} \in \mathbb{R}^K$,

$$\frac{1}{n} \left(Y - X\tilde{b} \right)^T \left(Y - X\tilde{b} \right) + \tau \tilde{b}^T \Lambda^2 \tilde{b} \tag{5}$$

where for (3) τ is chosen so that the constraint $\tilde{b}^T \Lambda \tilde{b} \leq B^2$ is satisfied. In this latter case, τ is necessarily random because the constraint needs to be satisfied in sample. Here the tilde in $\tilde{b}$ is used to remind us that in the matrix formulation, b is truncated to be a K dimensional vector, as all entries larger than K are zero by definition of $\mathcal{E}_K$. The solution is the usual ridge regression estimator $\tilde{b}^{(n,\tau)} := \left(X^T X + n\tau\Lambda^2 \right)^{-1} X^T Y$.

For problem (4), $\tau = \tau_n$ can go to zero in a controlled way. For problem (3), $\tau = \tau_{B,n} \geq 0$ must be chosen so that the constraint is satisfied. Such $\tau_{B,n}$ is non-zero if the constraint is binding, and zero otherwise.

All vectors are in $\mathbb{R}^\infty$, though only the first K elements might be non-zero. The exception is when we use a tilde, as in (5). For $b^{(n)}$ in (3), the ℓ_2 norm of $b^{(n)} - \varphi$ becomes $\left|b^{(n)} - \varphi\right|_2 = \left(\sum_{k=1}^K \left|b_k^{(n)} - \varphi_k\right|^2 + \sum_{k>K} |\varphi_k|^2\right)^{1/2}$

It is worth noting that the ellipsoid $\mathcal{E} \subset \ell_2$ is a RKHS generated by the kernel $C(k,l) = \sum_{v=1}^\infty \lambda_v^{-2} \delta_{v,k}\delta_{v,l}$ where $\delta_{v,l}$ is the Kronecker's delta, i.e. $\delta_{v,l} = 1$ if $v = l$ and zero otherwise. The inner product $\langle \cdot, \cdot \rangle_{\mathcal{E}}$ is defined to satisfy the reproducing kernel property $\langle C(\cdot, l), C(\cdot, k)\rangle_{\mathcal{E}} = C(k,l)$. Hence for $a, b \in \mathcal{E}$, $b_k = \langle b, C(\cdot, k)\rangle_{\mathcal{E}}$ and $\langle a, b\rangle_{\mathcal{E}} = \sum_{v=1}^\infty \lambda_v^2 a_v b_v$. The norm induced by the inner product is $|\cdot|_{\mathcal{E}}$ such that for any vector $b \in \mathbb{R}^\infty$, $|b|_{\mathcal{E}}^2 = \sum_{k=1}^\infty \lambda_k^2 b_k^2$. This norm dominates the ℓ_2 norm. The fact that $\mathcal{E}(1)$ is compact under the ℓ_2 norm is a consequence of the fact that $\mathcal{E}$ is a RKHS (Kuelbs, 1976, Lemma 2.1(iv) and example 2; see also Li and Linde, 1999) and sharp asymptotics can be derived by related means (Graf and Luschgy, 2004).

We shall not directly use this fact in the proofs. However, once we realize such compactness, it is not a surprise that it might be possible to estimate infinite AR processes under no restriction on the number of estimated coefficients except for being in $\mathcal{E}(B)$. We also establish convergence rates. Moreover, we want to clearly address the relation between constrained and penalized estimation.

The best approximation $\varphi^{(K)} \in \mathcal{E}_K$ to φ minimizes the population mean square error

$$\varphi^{(K)} = \arg \inf_{b \in \mathcal{E}_K} \mathbb{E}\left(Y_1 - \sum_{k=1}^\infty b_k Y_{1-k}\right)^2 \tag{6}$$

Theorem 1 *Suppose that Condition 1, and n, $K \to \infty$ hold.*

1. *(Consistency of Constrained Estimator) If $\varphi \in \mathcal{E}(B)$, then, for any $\epsilon \in (0, 2\eta - 1)$, $\left|b^{(n)} - \varphi\right|_2 = O_p\left(n^{-\frac{1}{2}\left(\frac{2\eta-\epsilon}{2\eta-\epsilon+1}\right)} + \left(K^\eta \ln^{1/2}(K)\right)^{-1}\right)$.*

2. *(Consistency of Penalized Estimator) Consider possibly random $\tau = \tau_n$ such that $\tau \to 0$ and $\tau n^{1/2} \to \infty$ in probability. Then, $\left|b^{(n,\tau)} - \varphi\right|_{\mathcal{E}} \to 0$ in probability. Hence, for any real B such that $|\varphi|_{\mathcal{E}} < B$, then, $\left|b^{(n,\tau)}\right|_{\mathcal{E}} < B$ with probability going to one.*

3. *(Equivalence of Estimators) If $\varphi \in \mathcal{E}(B)$, there is a random $\tau = \tau_{B,n}$ such that $\tau = O_p\left(n^{-1/2}\right)$, $b^{(n,\tau)} = b^{(n)}$.*

4. *(Consistency in $\mathcal{E}_K$) If $\tau \asymp n^{-1/4}K^{-\eta}$, then $\left|b^{(n,\tau)} - \varphi^{(K)}\right|_{\mathcal{E}} = O_p\left(n^{-1/4}K^\eta\right)$.*

5. *(Approximation Error in $\mathcal{E}_K$) We have that $\left|\varphi - \varphi^{(K)}\right|_{\mathcal{E}} = o\left(1\right)$. Moreover, suppose that the k^{th} entry φ_k in φ satisfies $\left|\varphi_k\right| \lesssim k^{-\nu}$ with $\nu > \left(2\eta + 1\right)/2$ for all k large enough. Then $\left|\varphi - \varphi^{(K)}\right|_{\mathcal{E}} = O\left(K^{(2\eta+1-2\nu)/2}\right)$.*

6. *(Difference Between Norms) Suppose that $\varphi \in \operatorname{int}\left(\mathcal{E}\left(B\right)\right)$ ($\operatorname{int}\left(\mathcal{E}\left(B\right)\right)$ is the interior of $\mathcal{E}\left(B\right)$). Then, we can find $K \to \infty$ and $\tau = O_p\left(n^{-1/2}\right)$ such that $\left|b^{(n,\tau)} - \varphi\right|_2 \to 0$ in probability, but $\left|b^{(n,\tau)} - \varphi\right|_{\mathcal{E}}$ does not converge to zero in probability.*

Point 1 establishes the convergence rate of (3) towards the true φ in terms of the exponent η (in Condition 1). This rate does not constrain the number of lags used as long as $\varphi \in \mathcal{E}\left(B\right)$. For the finite dimensional case we trivially recover the root-n convergence by letting $\eta \to \infty$.

Point 2 says that if we use the penalized estimation and the penalty does not go to zero too fast (i.e. strictly slower than in Point 1), (4) is consistent under the norm $\left|\cdot\right|_{\mathcal{E}}$. Moreover, with probability going to one, (4) will be contained in a ball in $\mathcal{E}$ that contains the true parameter.

Point 3 establishes the link between constrained and penalized estimation by finding the rate of decay of the ridge penalty so that (3) and (4) are the same.

Point 4 provides the rate of consistency of (6) for $\varphi^{(K)}$ under the RKHS norm. The latter parameter defines the closest AR model in $\mathcal{E}_K$ to the true infinite AR model.

Point 5 shows how close $\varphi^{(K)}$ is to φ under the RKHS norm for $K \to \infty$.

Point 6 establishes an additional insight between the convergence under the ℓ_2 norm and the RKHS norm in terms of the penalty. A "slowly convergent" penalty is necessary for convergence under $\left|\cdot\right|_{\mathcal{E}}$. Hence, this also shows that the constrained estimator (whose penalty is $\tau = \tau_{B,n} = O_p\left(n^{-1/2}\right)$ when $\varphi \in \mathcal{E}\left(B\right)$) cannot be consistent in the norm $\left|\cdot\right|_{\mathcal{E}}$ in general. This happens when choosing a rather large K that leads to a binding constraint for (3).

Combining Points 4 and 5 in Theorem 1, we have the following.

Corollary 1 *Suppose that Condition 1 holds, $K \asymp n^{\frac{1}{2(2\nu-1)}}$, $\tau \asymp n^{-1/4}K^{-\eta}$, and that the k^{th} entry φ_k in φ satisfies $\left|\varphi_k\right| \lesssim k^{-\nu}$ with $\nu > \left(2\eta + 1\right)/2$ for all k large enough. Then, $\left|b^{(n,\tau)} - \varphi\right|_{\mathcal{E}} = O_p\left(n^{-\frac{2\nu-(2\eta+1)}{4(2\nu-1)}}\right)$.*

Corollary 1 imposes additional restrictions in order to improve on the statement of Point 2 in Theorem 1, by giving rates of convergence. These rates are not tight as they require $K = o(n)$ unlike Point 2 in Theorem 1. However, they are useful in applications (e.g. Section 2.1.1).

Sieve estimators are often consistent under the sole condition that the number of components (here K) is of smaller order of magnitude than the sample size n. In Point 1 of Theorem 1, we have shown that this is not required. Recall that $N = n + K$ is the sample size. We can have $K = O(N)$ as long as $n \to \infty$. Of course, we require knowledge concerning the magnitude of the coefficients. Such knowledge is usually assumed in the literature in order to bound the approximation error.

In practice the fact that we allow $K = O(N)$ might sound irrelevant. However, the asymptotic results can be seen as suggesting that, once we set the constraint, the procedure used here can be more robust to lag choice. We show this in the simulation in Section 3.

2.1.1 Application to Optimal Forecasting and Universal Consistency

Define $X_t(a) = \sum_{k=1}^{\infty} a_k Y_{t-k}$ for any $a \in \mathbb{R}^{\infty}$. The expectation of Y_t conditioning on the infinite past $(Y_{t-s})_{s\geq 1}$ is $X_t(\varphi)$. As an application of Theorem 1 consider the following problem. Show that, as $n \to \infty$,

$$\sup_{t\in\mathcal{T}} \left|X_t(\varphi) - X_t\left(b^{(n,\tau)}\right)\right| \to 0$$

in probability where $\mathcal{T} = (0, \infty)$ or $(0, n)$ ($b^{(n,\tau)}$ in (4)). The above display is a weak form of universal consistency because the convergence is in probability rather than almost surely. We want $X_t\left(b^{(n,\tau)}\right)$ to be close to the conditional expectation of Y_t uniformly in $t \in \mathcal{T}$, which is even more general than considering a moving target. The norm $|\cdot|_{\mathcal{E}}$ is useful because the previous display can be written as

$$\sup_{t\in\mathcal{T}} \left|X_t\left(\varphi - b^{(n,\tau)}\right)\right| \lesssim \left|\varphi - b^{(n,\tau)}\right|_{\mathcal{E}} \sup_{t\in\mathcal{T}} \left(\sum_{k=1}^{\infty}\left(\frac{Y_{t-k}}{k^{\eta}}\right)^2\right)^{1/2}. \tag{7}$$

To obtain the inequality, we have multiplied and divided each term in the sum (on the l.h.s.) by λ_k and then used the Cauchy-Schwarz inequality and Condition 1 to set $\lambda_k^{-1} \lesssim k^{-\eta}$.

We have that $\left|\varphi - b^{(n,\tau)}\right|_{\mathcal{E}} = O_p(\epsilon_n)$ in probability, where $\epsilon_n \to 0$ at a rate which depends on Theorem 1. Then, if

$$\sup_{t\in\mathcal{T}}\left(\sum_{k=1}^{\infty}\left(\frac{Y_{t-k}}{k^{\eta}}\right)^{2}\right)^{1/2}=o_p\left(\epsilon_n^{-1}\right), \tag{8}$$

we have shown that (7) goes to zero in probability. The convergence of (7) to zero holds for a variety of processes and circumstances.

If $\mathcal{T}=(0,\infty)$ then (8) is almost surely finite if the random variables are bounded, and (7) goes to zero in probability using Point 2 in Theorem 1.

If $\mathcal{T}=(0,n)$, we can use the bound

$$\left(\mathbb{E}\sup_{t\in(0,n)}\sum_{k=1}^{\infty}Y_{t-k}^{2}k^{-2\eta}\right)^{1/2}\leq n^{1/(2p)}\sup_{t\in(0,n)}\left(\mathbb{E}\sum_{k=1}^{\infty}Y_{t-k}^{2p}k^{-2\eta p}\right)^{1/(2p)}$$

when the variables are $2p$ integrable (Lemma 2.2.2 in van der Vaart and Wellner, 2000). For any $2p$ integrable sequence $(Z_t)_{t\in\mathbb{Z}}$, this is a consequence of the following chain of inequalities

$$\mathbb{E}\left(\max_{t\leq n}|Z_t|^{2}\right)^{1/2}\leq\left(\mathbb{E}\max_{t\leq n}|Z_t|^{2p}\right)^{1/(2p)}\leq\left(\sum_{t=1}^{n}\mathbb{E}|Z_t|^{2p}\right)^{1/(2p)}\leq n^{1/(2p)}\left(\max_{t\leq n}\mathbb{E}|Z_t|^{2p}\right)^{1/(2p)}.$$

If p is such that $n^{1/(2p)}=o\left(\epsilon_n^{-1}\right)$, then the r.h.s. of (7) goes to zero in probability. If Y_t has moment generating function, the r.h.s. of the above display is actually $O\left(\sqrt{\ln n}\right)$ (Lemma 2.2.2 in van der Vaart and Wellner, 2000). Either way, to find ϵ_n we can use Corollary 1. Note that the argument is unchanged if $\mathcal{T}=(0,c_n)$ for any $c_n\asymp n$.

Theorem 1 can also be applied to the less ambitious problem: show that

$$\lim_{K\to\infty}\sup_{t\in\mathcal{T}}\left|X_t\left(\varphi^{(K)}\right)-X_t\left(b^{(n,\tau)}\right)\right|\to 0$$

in probability. In this case we want to forecast as well as the increasingly best approximation of the conditional expectation of Y_t, uniformly in $t\in\mathcal{T}$. Point 4 in Theorem 1 is suited for this problem.

2.2 Choice of B in Practice

The parameter B can be chosen to minimize some cross-validated prediction error estimate (beware of cross-validation in a time series context, e.g. Györfi et al., 1990, Burman and Nolan, 1992, Burman et al., 1994, for discussions and applicability). Alternatively, one can choose B to minimize a penalized loss function such as

$$\ln \hat{\sigma}_B^2 + \frac{2\mathrm{df}\left(B\right)}{n} \tag{9}$$

where $\mathrm{df}\left(B\right) = \mathrm{Trace}\left(\left(X^T X + \tau_{B,n} n \Lambda^2\right)^{-1} X^T X\right)$ using the notation in (5). Here, $\hat{\sigma}_B^2 = \hat{e}^T \hat{e}/n$ and $\hat{e}$ is the n dimensional vector of residuals $\hat{e} = Y - X\tilde{b}^{(n)}$. In particular, if the constraint is binding, $\tau_{B,n}$ is implicitly defined by

$$Y^T X \left(X^T X + \tau_{B,n} n \Lambda^2\right)^{-2} X^T Y = B^2, \tag{10}$$

otherwise it is zero. The procedure is the same as for lag selection using Akaike's information criterion (AIC) when the constraint is not binding; set $\tau_{B,n} = 0$ in the definition of $\mathrm{df}\left(B\right)$ to see this. Hence, the number of lags/parameters is replaced by $\mathrm{df}\left(B\right)$, which is the effective number of degrees of freedom implied by B (Hastie et al., 2009).

3 Numerical Example

Asymptotic results are of interest on their own, but it is also relevant to understand the scope of applicability in practice. As a benchmark, we use predictions based on an AR model estimate where the lag length is chosen using AIC.

3.1 Simulated True Model

One thousand data samples are simulated from (1). The sample size is $N = 1000$. A burn in of 20000 observations is used to reduce any dependence on the starting value. We simulate a testing sample of 1000 observations to approximate the mean square error (MSE). We consider different parametrizations of the ARFIMA model

$$Y_t = \sum_{k=1}^{K_0} \varphi_k Y_{t-k} + \left(1 - \mathrm{L}\right)^{-d} \left(\sum_{l=0}^{L} \theta_l \varepsilon_{t-l}\right) \tag{11}$$

The symbol L stands for the lag operator (not to be confused with the constant L): $\mathrm{L}\varepsilon_t = \varepsilon_{t-1}$. The MA polynomial is $\theta_l = (1 - 0.1l)$ with $L = 5$. The coefficient of fractional integration is $d \in \{0, 0.49\}$. We choose the errors in (1) i.i.d. t-distributed with $v \in \{2, 4, 30\}$ degrees of freedom. When $v \in \{2, 4\}$, the innovations do not satisfy Condition 1. In fact, for v equal to 2 and 4, the innovations have infinite variance and infinite kurtosis respectively. When $v = 30$, the innovations are essentially Gaussian. The coefficients of the autoregression are chosen to be $\varphi_k = \bar{\varphi}k^{-1/2}/\left(\sum_{k=1}^{K_0} k^{-1/2}\right)$, where $\bar{\varphi} \in \{0.75, 0.99\}$. A higher value for $\bar{\varphi}$ leads to a more persistent behaviour. By construction, for both values of $\bar{\varphi}$, the model appears to generate cycles because the roots of $1-\sum_{k=1}^{K_0} \varphi_k z^k = 0$ are outside the unit circle, but complex. We shall have different values for $K_0 \in \{10, 100, 1000\}$. The scalar K_0 is the unknown true lag length. The simulation design accounts for short and long memory infinite autoregressive processes.

Short Memory When $d = 0$ the model reduces to an invertible ARMA model of finite order. In this case, (11) has an infinite AR representation with asymptotically geometrically decaying coefficients. The latter claim follows from invertibility, and the fact that K_0 is finite. In consequence, the coefficients of the autoregression are in $\mathcal{E}$.

Long Memory Model When $d = 0.49$, the model is stationary, but exhibits long memory. In particular, from the binomial expansion, $(1 - \mathrm{L})^d = \sum_{j=1}^{\infty} \pi_j \mathrm{L}^j$ with nonzero π_j's (Brockwell and Davis, 1991, eq. 13.2.2 for details). Hence, the process is an AR(∞). It admits an MA(∞) representation with coefficients that are only square summable so that the autocorrelation function is not summable (Brockwell and Davis, 1991, Theorem 13.2.1). In this case, $\varphi \notin \mathcal{E}$ and an approximation error is incurred.

In practice, because of limitations in floating point arithmetic, the expansion of $(1 - \mathrm{L})^d$ is truncated after 100 terms. Hence, strictly speaking the process is an ARMA(K_0,100), but in finite samples, the behaviour is similar to a long memory process.

3.2 Estimation and Results

The parameter's estimates are obtained from (5) with $\lambda_k = k^\eta$ with $\eta \in \{0.501, 1\}$. This is to establish the sensitivity to the choice of λ in Condition 1. The benchmark is an AR model estimate with lag length chosen to minimize AIC. Denote the number of lags chosen using AIC by K_{AIC}. We compare this to a model estimated using more lags, but with coefficients constrained in $\mathcal{E}_K(B)$. In particular, $K = 2K_{AIC}$ and $4K_{AIC}$ with B chosen as outlined in Section 2.2. The goal is to verify whether the procedure is robust to lag choice. AIC is known to choose large models. We use even larger models, and verify whether we are able to obtain sensible results.

The results in Table 1 show the MSE of the constrained procedure over the MSE obtained by estimating an AR model with lag length selected by AIC. The numbers have been multiplied by 100. In the interest of space, only the results for $\bar{\varphi} = 0.75$ and $v = 30$ are reported in Table 1. The full set of results is essentially identical: the MSE increases for all procedures when $\bar{\varphi} = 0.99$ and $\nu \in \{2, 4\}$, but the ratio of the MSE's does not change.

These results show that the procedure is robust against lag choice. This becomes evident in the long memory case. The larger model ($4K_{AIC}$) leads to relatively better performance especially when the true model exhibits persistency ($d = 0.49$). We also deduce that, for the given simulation design, the choice of exponent η does not make a difference.

Table 1: Simulation Results. The true model is as in (11) with number of true AR coefficients equal to K_0 and AR coefficients satisfying $\varphi_k = \bar{\varphi} k^{-1/2} / \left(\sum_{k=1}^{K_0} k^{-1/2} \right)$, where $\bar{\varphi} = 0.75$ and t-distributed innovations with degrees of freedom $v = 30$. Entries denote the MSE of the constrained procedure over the MSE obtained by estimating an AR model with lag length K_{AIC} selected by AIC. Numbers are multiplied by 100. A number less than 100 favours the constrained procedure over AIC. The MSE in the numerator is computed using lag lengths $2K_{AIC}$ and $4K_{AIC}$ and B chosen as described in Section 2.2. The ellipsoid is defined by $\lambda_k = k^\eta$ for $k > 0$.

$K_0 =$	10		100		1000	
	$2K_{AIC}$	$4K_{AIC}$	$2K_{AIC}$	$4K_{AIC}$	$2K_{AIC}$	$4K_{AIC}$
	Short Memory: $d = 0$					
$\eta = 0.501$	97	95	95	92	94	91
$\eta = 1$	97	95	95	92	94	91
	Long Memory: $d = 0.49$					

$\eta = 0.501$	92	87	93	88	94	88
$\eta = 1$	92	87	93	88	94	88

4 Further Remarks

It is simple to impose linear restrictions on the coefficients of either the constrained or penalized estimator. A natural example is positivity. This is the case if we wish to estimate ARCH models of large orders. Under ARCH restrictions, the squared returns follow an AR process. The estimator does not have a closed form expression, but it is just the solution of a quadratic programming problem. Another extension pertains to vector autoregressive processes

$$Y_t = \sum_{k=1}^{\infty} \Phi_k Y_{t-k} + \varepsilon_t \tag{12}$$

where now the variables and innovations are L dimensional vectors and we use the capital Φ_k to stress the multivariate framework, where Φ_k is an $L \times L$ matrix. Again, we can restrict $\mathcal{E}$ in a suitable way. For example, Φ_k can be restricted to be lower triangular. This restriction has a variety of implications going from Granger causality to exogeneity and it is of much interest in econometrics (e.g., Sims, 1980). For fixed L, all the results in this paper apply to this problem as well, with obvious changes if we modify the constraint to $\sum_{k=1}^{\infty} |\Phi_k|^2 \lambda_k^2 \leq B$ where $|\Phi_k|$ is any matrix norm, e.g., Frobenius: $|\Phi_k| = \sqrt{Trace\left(\Phi_k^T \Phi_k\right)}$.

An extension, which does not follow directly from the results derived here, is to consider the case where $L \to \infty$. This is the problem where we have a large cross-section (L is the dimensional of the vector Y_t in (12)). In this case, the constraint cannot use an arbitrary matrix norm (norms are not equivalent in infinite dimensional spaces). Results in Lutz and Bühlmann (2006) together with the ones derived here can provide initial guidance on how to tackle this problem in the future.

Finally, the paper has not considered the asymptotic distribution of the constrained and penalized estimators. For finite dimensional regression problems with i.i.d. observations, the distribution of constrained estimators and the penalized estimators is well known (Geyer, 1994, Fu and Knight, 2000). The results in Section 5 can be used to extend those results to our autoregressive problem when K is bounded. When K diverges to infinity, which is the focus of this paper, to the author's knowledge, there are no available results directly applicable to the estimators in (3) and (4).

5 Proofs

Lemma 1 is standard, but a proof is provided for convenience.

Proof. [Lemma 1] A stationary infinite AR process with absolutely summable AR coefficients has an infinite MA representation with absolutely summable coefficient and it is invertible (Lemma 2.1 in Bühlmann, 1995). Hence, there are coefficients ψ_s such that $Y_t = \sum_{s=0}^{\infty} \psi_s \varepsilon_{t-s}$ and

$$\sum_{k=1}^{\infty} |\mathbb{E}Y_t Y_{t-k}| \leq \sigma^2 \sum_{k=1}^{\infty} \sum_{s=0}^{\infty} |\psi_{s+k}| \, |\psi_s| < \infty.$$

This means that the autocovariance function is absolutely summable. The moment bound follows from the infinite MA representation and the bound on the fourth moment of the innovations. The process is ergodic. This follows from the fact that it is integrable, and it is a filter of i.i.d. observations whose invariant sets are trivial by Kolmogorov 0-1 law. ■

5.1 Proof of Theorem 1

We divide the proof into two parts. The first only considers result under the ℓ_2 norm (Theorem 1, Point 1). The other is concerned with convergence results under the RKHS norm and the relation between the penalized and constrained estimator (Theorem 1, Points 2-6).

5.1.1 Consistency Under the ℓ_2 Norm (Point 1 in Theorem 1)

This section is concerned with the proof of Point 1 in Theorem 1. Few lemmas are needed for the proof. Throughout, we shall use the notation $X_t(a) = \sum_{k=1}^{\infty} a_k Y_{t-k}$ for any $a \in \mathbb{R}^{\infty}$. The proof of Point 1 in Theorem 1 can be found at the end of this section.

Lemma 3 *For $\rho := (2\eta + 1)/2 > 1$ ($\eta > 1/2$ as in Condition 1) and real constants w_k, there is a finite constant c_B - that depends on B - such that $\sup_{b \in \mathcal{E}(B)} \left|\sum_{k=1}^{\infty} b_k w_k\right| \leq c_B \sum_{k=1}^{\infty} k^{-\rho} |w_k|$, and similarly, for real constants $w_{k,l}$, $\sup_{b \in \mathcal{E}(B)} \left|\sum_{k,l=1}^{\infty} b_k b_l w_{lk}\right| \leq c_B^2 \sum_{k,l=1}^{\infty} k^{-\rho} l^{-\rho} |w_{kl}|$.*

Proof. Note that $\left|\sum_{k=1}^{\infty} b_k w_k\right| \leq \sum_{k,l=1}^{\infty} \frac{|b_k|}{k^{-\rho}} k^{-\rho} |w_k|$. Given that $b \in \mathcal{E}(B)$, $|b_k| \leq c_B k^{-\rho}$, by Lemma 2. This implies that the previous quantity is bounded by $c_B \sum_{k=1}^{\infty} k^{-\rho} |w_k|$. The same argument proves the second statement in the lemma ■

The coefficients w_{kl} in the lemma above will be partial sums of cross products of Y_t's, which we bound using the following.

Lemma 4 *Under Condition 1,*

$$\sup_{n,k,l>0} \mathbb{E} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (1-\mathbb{E}) Y_{t-k} Y_{t-l} \right|^2 < \infty.$$

Proof. From the proof of Lemma 1, there are absolutely summable coefficients ψ_u, such that $Y_t = \sum_{u=0}^{\infty} \psi_u \varepsilon_{t-u}$. For ease of notation in what follows suppose that the i.i.d. innovations have variance one and the MA coefficients are non-negative.

For any square integrable stationary sequence of mean zero random variables $(Z_t)_{t\in\mathbb{Z}}$,

$$\mathbb{E} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} Z_t \right|^2 \leq Var(Z_t) + 2 \sum_{s=1}^{n-1} Cov(Z_t, Z_{t+s}) \leq 2 \sum_{s=0}^{n-1} Cov(Z_t, Z_s)$$

for any integer t, by stationarity. By this remark, setting $Z_t = (1-\mathbb{E}) Y_{t-k} Y_{t-l}$, deduce that

$$\mathbb{E} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (1-\mathbb{E}) Y_{t-k} Y_{t-l} \right|^2 \leq 2 \sum_{s=0}^{n-1} \mathbb{E} \left[(1-\mathbb{E}) Y_{t-k} Y_{t-l} (1-\mathbb{E}) Y_{t-s-k} Y_{t-s-l} \right],$$

where the r.h.s. holds for $t \in \mathbb{Z}$. If we showed that

$$\mathbb{E} \left[(1-\mathbb{E}) Y_{t-k} Y_{t-l} (1-\mathbb{E}) Y_{t-s-k} Y_{t-s-l} \right] \lesssim \psi_s$$

the result would follow by summability of the coefficients. To show the above, with no loss of generality, by symmetry, consider only the case $l \geq k$. This implies that

$$\begin{aligned}
& \mathbb{E}\left[(1-\mathbb{E}) Y_{t-k} Y_{t-l} (1-\mathbb{E}) Y_{t-s-k} Y_{t-s-l}\right] \\
= \; & Cov\left(Y_{t-k} Y_{t-l}, Y_{t-s-k} Y_{t-s-l}\right) \\
= \; & \mathbb{E} \sum_{u_1=0}^{\infty} \sum_{u_2=0}^{\infty} \psi_{u_1} \psi_{u_2} (1-\mathbb{E}) \varepsilon_{t-k-u_1} \varepsilon_{t-l-u_2} \\
& \times \sum_{u_3=0}^{\infty} \sum_{u_4=0}^{\infty} \psi_{u_3} \psi_{u_4} (1-\mathbb{E}) \varepsilon_{t-s-k-u_3} \varepsilon_{t-s-l-u_4}.
\end{aligned}$$

The above is equal to

$$\sum_{u_1=0}^{\infty} \sum_{u_2=0}^{\infty} \sum_{u_3=0}^{\infty} \sum_{u_4=0}^{\infty} \psi_{u_1} \psi_{u_2} \psi_{u_3} \psi_{u_4} Cov\left(\varepsilon_{t-k-u_1} \varepsilon_{t-l-u_2}, \varepsilon_{t-s-k-u_3} \varepsilon_{t-s-l-u_4}\right).$$

By the i.i.d. condition on the innovations, the covariance is zero if the indices are not constrained in the following sets $\{k+u_1 = l+u_2,\ k+u_3 = l+u_4\}$, $\{u_1 = u_3+s,\ u_2 = u_4+s\}$, $\{k+u_1 = l+u_4+s,\ l+u_2 = k+u_3+s\}$. Hence, we can consider summation with indices in these sets only. Splitting the sum according to the above index sets, we have respectively,

$$I = \sum_{u=0}^{\infty} \sum_{v=0}^{\infty} \psi_{u+l-k} \psi_u \psi_{v+l-k} \psi_v Cov\left(\varepsilon_0^2, \varepsilon_{u-(s+v)}^2\right),$$

$$II = \sum_{u=0}^{\infty} \sum_{v=0}^{\infty} \psi_{u+s} \psi_{v+s} \psi_u \psi_v \mathbb{E} \varepsilon_0^2 \varepsilon_{(u-v)+(k-l)}^2,$$

$$III = \sum_{u=0}^{\infty} \sum_{v=0}^{\infty} \psi_{u+s+(l-k)} \psi_{v+s+(k-l)} \psi_u \psi_v \mathbb{E} \varepsilon_0^2 \varepsilon_{(u-v-s)+(k-l)}^2.$$

From the fact that the innovations are i.i.d., $Cov\left(\varepsilon_0^2, \varepsilon_{u-(s+v)}^2\right) = 0$ unless $u = (s+v)$. Hence, we further constrain the sum to deduce that

$$I \lesssim \sum_{u=0}^{\infty} \sum_{v=0}^{\infty} \psi_{u+l-k} \psi_u \psi_{v+l-k} \psi_v 1_{\{u-v=s\}}$$

where $1_{\{\cdot\}}$ is the indicator function, which is one if the argument is true and zero otherwise. Substituting $u = v+s$ in the indices containing u, the r.h.s. of the above display is equal to

$$\sum_{v=0}^{\infty}\psi_{v+s+l-k}\psi_{v+s}\psi_{v+(l-k)}\psi_{v}.$$

Now note that the coefficients are asymptotically decreasing and that $l\geq k$, so that $\psi_{v+s+l-k}\lesssim\psi_{v+s}\lesssim\psi_{s}$ and $\psi_{v+(l-k)}\lesssim\psi_{v}$. Hence replacing these in the previous display deduce that

$$I\lesssim\sum_{v=0}^{\infty}\psi_{v+s}^{2}\psi_{v}^{2}\lesssim\psi_{s}^{2}\sum_{v=0}^{\infty}\psi_{v}^{2}\lesssim\psi_{s}^{2}$$

by summability of the coefficients. By finiteness of the fourth moment of the innovations,

$$II\lesssim\sum_{u=0}^{\infty}\sum_{v=0}^{\infty}\psi_{u+s}\psi_{v+s}\psi_{u}\psi_{v}=\left(\sum_{u=0}^{\infty}\psi_{u}\psi_{u+s}\right)^{2}.$$

Using again the fact that the coefficients are asymptotically decreasing ($\psi_{u+s}\lesssim\psi_{s}$), the above display is bounded by a constant multiple of $\psi_{s}^{2}\left(\sum_{u=0}^{\infty}\psi_{u}\right)^{2}\lesssim\psi_{s}^{2}$. Finally, by the moment bound, and the same arguments as before,

$$III\lesssim\sum_{u=0}^{\infty}\sum_{v=0}^{\infty}\psi_{u}\psi_{v}\psi_{u+s+(l-k)}\psi_{v+s+(k-l)}\leq\psi_{s}\left(\sum_{u=0}^{\infty}\sum_{v=0}^{\infty}\psi_{v}\psi_{u}\right)\lesssim\psi_{s}.$$

The second inequality used the fact that $\psi_{v+s+(k-l)}\lesssim 1$. The bounds do not depend on k,l beyond the fact that $l\geq k$. Repeating the argument for $k>l$, the result follows. ■

Lemma 4 will be used to bound quantities such as the following

$$\begin{aligned}&\mathbb{E}\left|\sum_{k,l=1}^{\infty}k^{-(2\eta+1)/2}l^{-(2\eta+1)/2}\frac{1}{n}\sum_{t=1}^{n}\left(1-\mathbb{E}\right)Y_{t-k}Y_{t-l}\right|\\ \leq&\sum_{k,l=1}^{\infty}k^{-(2\eta+1)/2}l^{-(2\eta+1)/2}\mathbb{E}\left|\frac{1}{n}\sum_{t=1}^{n}\left(1-\mathbb{E}\right)Y_{t-k}Y_{t-l}\right|\\ \lesssim&\frac{1}{\sqrt{n}}\max_{k,l>0}\mathbb{E}\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(1-\mathbb{E}\right)Y_{t-k}Y_{t-l}\right|,\end{aligned}$$

where the second inequality follows because $(2\eta+1)/2>1$. Then, by Lemma 4 the expectation is finite because $\mathbb{E}\left|\cdot\right|\leq\left(\mathbb{E}\left|\cdot\right|^{2}\right)^{1/2}$ and it is independent of k,l by stationarity. In consequence the display is $O_{p}\left(n^{-1/2}\right)$ because convergence in L_{1} implies convergence in probability.

To establish convergence rates we need two stochastic equicontinuity results.

Lemma 5 *Under Condition 1, for any $\epsilon \in (0, 2\eta - 1)$,*

$$\mathbb{E} \sup_{a,b \in \mathcal{E}(2B), |b|_2 \leq \delta} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (1 - \mathbb{E}) X_t(b) X_t(a) \right| \lesssim \delta^{\frac{2\eta - \epsilon - 1}{2\eta - \epsilon}}, \tag{13}$$

where η is the exponent in Condition 1.

Proof. By the triangle inequality, the l.h.s. in (13) is bounded by

$$\mathbb{E} \sup_{a,b \in \mathcal{E}(2B), |b|_2 \leq \delta} \sum_{l=1}^{\infty} |a_l| \sum_{k=1}^{\infty} |b_k| \left| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (1 - \mathbb{E}) Y_{t-k} Y_{t-l} \right|.$$

By Lemma 3, there is a $\rho > 1$ such that the above is bounded by a constant multiple of

$$\begin{aligned} &\sum_{l=1}^{\infty} l^{-\rho} \mathbb{E} \sup_{b \in \mathcal{E}(2B), |b|_2 \leq \delta} \sum_{k=1}^{\infty} |b_k| \left| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (1 - \mathbb{E}) Y_{t-k} Y_{t-l} \right| \\ \lesssim\ & \sup_{l>0} \mathbb{E} \sup_{b \in \mathcal{E}(2B), |b|_2 \leq \delta} \sum_{k=1}^{\infty} |b_k| \left| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (1 - \mathbb{E}) Y_{t-k} Y_{t-l} \right| \end{aligned}$$

by summability of $l^{-\rho}$. For any positive V, the above display can be written as

$$\sup_{l>0} \mathbb{E} \sup_{b \in \mathcal{E}(2B), |b|_2 \leq \delta} \left(\sum_{k \leq V} + \sum_{k > V} \right) |b_k| \left| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (1 - \mathbb{E}) Y_{t-k} Y_{t-l} \right|. \tag{14}$$

We shall bound the two sums separately. By the Cauchy-Schwarz inequality, the first sum is bounded by

$$\sqrt{\sup_{l>0} \sup_{|b|_2 \leq \delta} \sum_{k \leq V} b_k^2 \sum_{k \leq V} \mathbb{E} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (1 - \mathbb{E}) Y_{t-k} Y_{t-l} \right|^2} \lesssim \delta \sqrt{V}, \tag{15}$$

where the inequality uses Lemma 4 and $|b|_2 \leq \delta$. Given that V has been fixed, multiplying and dividing each term by $k^{1+\epsilon}$ with $\epsilon \in (0, 2\eta - 1)$, and then using the Cauchy-Schwarz inequality, the second sum in (14) is bounded by

$$\sqrt{\left(\sup_{b \in \mathcal{E}(2B)} \sum_{k > V} b_k^2 k^{1+\epsilon} \right) \left(\sup_{l>0} \sum_{k > V} k^{-(1+\epsilon)} \mathbb{E} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (1 - \mathbb{E}) Y_{t-k} Y_{t-l} \right|^2 \right)}.$$

Use the the fact that $k^{-(1+\epsilon)}$ is summable, together with Lemma 4, to deduce that the second term in the product above is bounded by a constant. Hence the above is bounded by a constant multiple of

$$\sqrt{\left(\sup_{b\in\mathcal{E}(2B)}\sum_{k>V}b_k^2k^{1+\epsilon}\right)}\lesssim\sqrt{V^{1+\epsilon}\lambda_V^{-2}\left(\sup_{b\in\mathcal{E}(2B)}\sum_{k>V}b_k^2\lambda_k^2\right)}.$$

To deduce the inequality, we multiplied and divided each addend by λ_k^2, and used the fact that $k^{1+\epsilon}\lambda_k^{-2}\asymp k^{\epsilon-(2\eta-1)}\leq V^{\epsilon-(2\eta-1)}$ is decreasing in $k>V$. The above display is then bounded by a constant multiple of $V^{(1+\epsilon-2\eta)/2}$. This together with the bound in (15) show that (14) is bounded above by a constant multiple of $\delta\sqrt{V}+V^{(1+\epsilon-2\eta)/2}\lesssim\delta^{\frac{2\eta-\epsilon-1}{2\eta-\epsilon}}$ when we choose $V=\delta^{-\frac{2}{2\eta-\epsilon}}$. This proves the bound in the lemma. ■

Lemma 6 *Under Condition 1, for any $\epsilon\in(0,2\eta-1)$,*

$$\mathbb{E}\sup_{b\in\mathcal{E}(2B),|b|_2\leq\delta}\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_tX_t(b)\right|\lesssim\delta^{\frac{2\eta-\epsilon-1}{2\eta-\epsilon}},$$

where η is the exponent in Condition 1.

Proof. By linearity and the triangle inequality,

$$\begin{aligned}&\mathbb{E}\sup_{b\in\mathcal{E}(2B),|b|_2\leq\delta}\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_tX_t(b)\right|\\ \leq\ &\mathbb{E}\sup_{b\in\mathcal{E}(2B),|b|_2\leq\delta}\sum_{k=1}^{\infty}|b_k|\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_tY_{t-k}\right|.\end{aligned}$$

Note that

$$\sup_{k>0}\mathbb{E}\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_tY_{t-k}\right|^2\leq\sigma^2\gamma(0),$$

where $\gamma(0)$ is the autocovariance function of $(Y_t)_{t\in\mathbb{Z}}$ at zero. We can then proceed exactly as in the proof of Lemma 5 to deduce the result. ■

Proof of Theorem 1 Point 1. Define the empirical loss function

$$L_n(b):=\frac{1}{n}\sum_{t=1}^{n}\left(Y_t-\sum_{k=1}^{\infty}b_kY_{t-k}\right)^2$$

where $b \in \mathcal{E}$. When $b \in \mathcal{E}_K(B)$ the sum inside the parenthesis only runs from 1 to K. The population loss is

$$L(b) := \mathbb{E}\left[X_1(\varphi - b)\right]^2. \tag{16}$$

Define $\beta = \beta^{(K)} \in \mathbb{R}^\infty$ such that its first K entries are as in φ and the remaining are all zero. The consistency proof is standard (van der Vaart and Wellner, 2000, Theorem 3.2.5) once we show the following:

$$L(b) - L(\beta) \gtrsim |b - \beta|_2^2, \tag{17}$$

$$\mathbb{E} \sup_{b \in \mathcal{E}_K(B): |b-\beta|_2 \leq \delta} \left|\left[L_n(b) - L(b)\right] - \left[L_n(\beta) - L(\beta)\right]\right| \lesssim \frac{\delta^\alpha}{\sqrt{n}}, \tag{18}$$

for some $\alpha \in (0,2)$. Then, for any sequence $r_n \to \infty$ satisfying $r_n^{2-\alpha} \lesssim \sqrt{n}$, $L_n(b_n) \leq L_n(\beta) + O_p(r_n^{-2})$ and $|\varphi - \beta|_2 \lesssim r_n^{-1}$, we have that $\left|b^{(n)} - \varphi\right|_2 = O_p(r_n^{-1})$. We can then choose $r_n^2 = n^{1/(2-\alpha)}$.

At first we verify (17). At the end of the proof we shall show that we can restrict attention to b such that

$$L(b) - L(\beta) \gtrsim \sum_{k,l=1}^{\infty} (b_k - \beta_k)(b_l - \beta_l)\gamma(k-l), \tag{19}$$

where $\gamma(k)$ is the autocovariance function (ACF) of $(Y_t)_{t\in\mathbb{Z}}$. The estimator is uniquely identified if the matrix, say Γ, with (k,l) entry equal to $\gamma(k-l)$, is strictly positive definite with smallest eigenvalue $\theta_{\min} > 0$ (see remarks after Lemma 2.2. in Kreiss et al., 2011). This is the case if the spectral density of $(Y_t)_{t\in\mathbb{Z}}$, say $g(\omega)$, is bounded away from zero. The spectral density of the AR model (1) is given by $g(\omega) = (2\pi)^{-1}\sigma^2/\varphi(\omega)$, where $\varphi(\omega) = \left|\sum_{k=0}^{\infty}\varphi_k e^{-ik\omega}\right|^2$ with $\varphi_0 := 1$. Noting that by Condition 1, $\varphi(\omega) = \left|\sum_{k=0}^{\infty}\varphi_k e^{-ik\omega}\right|^2 \leq \left(\sum_{k=0}^{\infty}|\varphi_k|\right)^2 < \infty$, deduce that the eigenvalues of Γ are bounded away from zero. Hence,

$$L(b) - L(\beta) \geq \theta_{\min}^{-1}\sum_{k=1}^{\infty}(b_k - \beta_k)^2 = |b - \beta|_2^2, \tag{20}$$

and (17) holds.

Using the notation $Y_t = X_t(\varphi) + \varepsilon_t$, the empirical loss is equal to

$$L_n(b) = \frac{1}{n}\sum_{t=1}^{n}\left[\varepsilon_t^2 + X_t^2(\varphi - b) + 2\varepsilon_t X_t(\varphi - b)\right].$$

This implies that

$$\begin{aligned}&(L_n(b) - L(b)) - (L_n(\beta) - L(\beta))\\ =&\frac{1}{n}\sum_{t=1}^{n}\left[2\varepsilon_t X_t(\beta - b) + (1-\mathbb{E})\left(X_t^2(b-\varphi) - X_t^2(\beta-\varphi)\right)\right].\end{aligned}$$

To verify (18), we need to bound the above uniformly in $b \in \mathcal{E}(B)$ such that $|b-\beta|_2 \le \delta$. To this end, apply Lemma 6 to the first term on the r.h.s. to find that the uniform bound is a constant multiple of $n^{-1/2}\delta^{\frac{2\eta-\epsilon-1}{2\eta-\epsilon}}$ for any $\epsilon \in (0, 2\eta - 1)$. By basic algebraic manipulations, the second term on the r.h.s. of the display is

$$\begin{aligned}&(1-\mathbb{E})\left(X_t^2(b-\varphi) - X_t^2(\beta-\varphi)\right)\\ =\;& \frac{1}{n}\sum_{t=1}^{n}(1-\mathbb{E})Y_{t-k}Y_{t-l}(b_k-\beta_k)(b_l-\varphi_l)\\ &+\frac{1}{n}\sum_{t=1}^{n}(1-\mathbb{E})Y_{t-k}Y_{t-l}(\beta_k-\varphi_k)(b_l-\beta_l).\end{aligned}$$

The equality follows adding and subtracting $\frac{1}{n}\sum_{t=1}^{n}(1-\mathbb{E})Y_{t-k}Y_{t-l}(\beta_k-\varphi_k)(b_l-\varphi_l)$. Note that both $\varphi - b$ and $\beta - \varphi$ are in $\mathcal{E}(2B)$. We apply Lemma 5 to deduce that each term on the r.h.s. of the above display is uniformly bounded in L_1 by a constant multiple of $n^{-1/2}\delta^{\frac{2\eta-\epsilon-1}{2\eta-\epsilon}}$ for any $\epsilon \in (0, 2\eta-1)$ when $|b-\beta|_2 \le \delta$. Hence (18) is verified with $\alpha = \frac{2\eta-\epsilon-1}{2\eta-\epsilon}$. When we are only interested in a finite dimensional model, we can take $\eta \to \infty$ to deduce that $\alpha = 1$, which is the parametric rate.

To find r_n note that

$$L_n\left(b^{(n)}\right) - L_n(\beta) \le L_n(b_n) - \inf_{b\in\mathcal{E}_K(B)} L_n(b) = 0.$$

Also, $|\varphi-\beta|_2 = \left(\sum_{k>K}|\varphi_k|^2\right)^{1/2} = O\left(K^{-\eta}/\ln^{1/2}(K)\right)$. To see this, use Lemma 2 and bound the sum by the integral

$$\sum_{k>K} |\varphi_k|^2 \lesssim \int_K^\infty \frac{x^{-(2\eta+1)}}{\ln(1+x)} dx \leq \frac{1}{\ln(1+K)} \int_K^\infty x^{-(2\eta+1)} dx$$

using the fact that $\ln(1+x)$ is monotonically increasing, in the last inequality. Finally, integrate to derive the bound. Hence we deduce that $r_n^{-1} = n^{-\frac{1}{2}\left(\frac{2\eta-\epsilon}{2\eta-\epsilon+1}\right)} + \left(K^{-\eta}/\ln^{1/2}(K)\right)$ as stated in Point 1 of the theorem.

It remains to show that (19) holds true. We recall a result from van der Vaart and Wellner (2000, Problem 3.4.5) that says that for any elements x, y, z in a normed space with norm $|\cdot|$, we have that $|z-y|^2 + |z-y|^2 \geq (1-2/c)|x-y|^2$ whenever $|x-y| \geq c|z-y|$ for some $c \geq 2$. We apply this to elements in $\mathbb{R}^\infty$ with norm $|a|_\gamma = \sqrt{\sum_{k,l=1}^\infty a_k a_l \gamma(k-l)}$, $a \in \mathbb{R}^\infty$, where $\gamma(k)$ is the autocovariance function of the AR process at k. This is a norm because the minimum eigenvalue $\theta_{\min}$ of the matrix Γ defined above is strictly positive. By definition of the population loss in (16), $L(\beta) = |\varphi - \beta|_\gamma^2$. Recall that β has the first K entries as φ and the remaining equal to zero. As $K \to \infty$, we have shown that $|\varphi - \beta|_\gamma^2 = O\left(K^{-\eta}/\ln^{1/2}(K)\right)$. Hence, we suppose that eventually $L(b) \geq 2^{-1}L(\beta)$. If this were not the case, $|\varphi-\beta|_\gamma^2 > 2|\varphi - b|_\gamma^2$ and we could bound $2\left|\varphi - b^{(n)}\right|_2^2$ by $|\varphi-\beta|_2^2$ in virtue of the fact that $|a|_\gamma^2 \geq \theta_{\min}|a|_2^2$. In this case, the proof of Point 1 in the theorem would be trivial. Hence, we can apply the result stated above and deduce (19). This concludes the proof of Point 1 i Theorem 1.

5.1.2 Consistency Under the RKHS Norm (Points 2-6 in Theorem 1)

The proof depends on a few preliminary lemmas. The proof of Points 2-6 in Theorem 1 can be found at the end of this section.

Let $\varphi^{(\tau)} := \varphi^{(K,\tau)} \in \mathcal{E}_K$ be the penalized population estimator

$$\varphi^{(\tau)} = \arg\inf_{b\in\mathcal{E}_K} \mathbb{E}X_1^2(b-\varphi) + \tau|b|_{\mathcal{E}}^2. \tag{21}$$

The following can be deduced from Theorem 5.9 in Steinwart and Christmann (2008, eq. 5.14). The proof is given, as the context might seem different at first sight.

Lemma 7 *Suppose Condition 1. For arbitrary but fixed $\tau > 0$, consider $b^{(n,\tau)}$ and $\varphi^{(\tau)}$*

in (4) and (21) with K possibly diverging to infinity. Then,

$$\left|b^{(n,\tau)} - \varphi^{(\tau)}\right|_{\mathcal{E}} \leq \sqrt{\sum_{k=1}^{K} \frac{1}{\tau^2 \lambda_k^2} \left(\frac{1}{n} \sum_{t=1}^{n} (1 - \mathbb{E}) \left(Y_t - X_t \left(\varphi^{(\tau)}\right)\right) Y_{t-k}\right)^2},$$

where $b_k^{(n,\tau)}$ is the k^{th} entry in $b^{(n,\tau)}$, and similarly for $\varphi^{(\tau)}$.

Proof. By convexity of the square error loss, differentiating $\left(Y_t - X_t\left(b^{(n,\tau)}\right)\right)^2$ w.r.t. $b^{(n,\tau)}$ around $\varphi^{(\tau)}$ and rearranging,

$$\frac{2}{n}\sum_{t=1}^{n}\left(Y_t - X_t\left(\varphi^{(\tau)}\right)\right)\left(X_t\left(b^{(n,\tau)}\right) - X_t\left(\varphi^{(\tau)}\right)\right) \leq \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - X_t\left(b^{(n,\tau)}\right)\right)^2 - \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - X_t\left(\varphi^{(\tau)}\right)\right)^2.$$

Using the identity $2(x-y)y + (x-y)^2 = x^2 - y^2$ for any real x, y, we have that

$$2\tau\sum_{k=1}^{\infty}\lambda_k^2\left(b_k^{(n,\tau)} - \varphi_k^{(\tau)}\right)\varphi_k^{(\tau)} + \tau\sum_{k=1}^{\infty}\lambda_k^2\left(b_k^{(n,\tau)} - \varphi_k^{(\tau)}\right)^2 = \tau\sum_{k=1}^{\infty}\lambda_k^2\left|b_k^{(n,\tau)}\right|^2 - \tau\sum_{k=1}^{\infty}\lambda_k^2\left|\varphi_k^{(\tau)}\right|^2.$$

The above two displays imply that

$$\begin{aligned}
&\frac{2}{n}\sum_{t=1}^{n}\left(Y_t - X_t\left(\varphi^{(\tau)}\right)\right)\left(X_t\left(b^{(n,\tau)}\right) - X_t\left(\varphi^{(\tau)}\right)\right) \\
&+2\tau\sum_{k=1}^{\infty}\lambda_k^2\left(b_k^{(n,\tau)} - \varphi_k^{(\tau)}\right)\varphi_k^{(\tau)} + \tau\sum_{k=1}^{\infty}\lambda_k^2\left(b_k^{(n,\tau)} - \varphi_k^{(\tau)}\right)^2 \\
\leq\ & \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - X_t\left(b^{(n,\tau)}\right)\right)^2 + \tau\sum_{k=1}^{\infty}\lambda_k^2\left|b_k^{(n,\tau)}\right|^2 - \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - X_t\left(\varphi^{(\tau)}\right)\right)^2 - \tau\sum_{k=1}^{\infty}\lambda_k^2\left|\varphi_k^{(\tau)}\right|^2 \leq 0
\end{aligned}$$

where the most r.h.s. inequality follows because $b^{(n,\tau)}$ minimizes the empirical penalized risk. The first order conditions for $\varphi^{(\tau)}$ read

$$\varphi_k^{(\tau)} = -\frac{1}{\tau\lambda_k^2}\mathbb{E}\left(Y_t - X_t\left(\varphi^{(\tau)}\right)\right)Y_{t-k} \tag{22}$$

for $k \geq 1$. Substituting this in the l.h.s. of the previous display,

$$\begin{aligned}&\frac{2}{n}\sum_{t=1}^{n}\left(Y_t - X_t\left(\varphi^{(\tau)}\right)\right)\left(X_t\left(b^{(n,\tau)}\right) - X_t\left(\varphi^{(\tau)}\right)\right)\\ &-2\mathbb{E}\left(Y_t - X_t\left(\varphi^{(\tau)}\right)\right)\sum_{k=1}^{K}\left(b_k^{(n,\tau)} - \varphi_k^{(\tau)}\right)Y_{t-k} + \tau\sum_{k=1}^{K}\lambda_k^2\left(b_k^{(n,\tau)} - \varphi_k^{(\tau)}\right)^2 \leq 0.\end{aligned}$$

Rearranging and using the definition of $X_t\left(b^{(n,\tau)} - \varphi^{(\tau)}\right)$, deduce that

$$\begin{aligned}&\tau\sum_{k=1}^{K}\lambda_k^2\left(b_k^{(n,\tau)} - \varphi_k^{(\tau)}\right)^2\\ \leq\;& \frac{2}{n}\sum_{t=1}^{n}\left(\mathbb{E}-1\right)\left(Y_t - X_t\left(\varphi^{(\tau)}\right)\right)\sum_{k=1}^{K}\left(b_k^{(n,\tau)} - \varphi_k^{(\tau)}\right)Y_{t-k}\\ \leq\;& \sqrt{\sum_{k=1}^{K}\frac{2}{\lambda_k^2}\left(\frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}-1\right)\left(Y_t - X_t\left(\varphi^{(\tau)}\right)\right)Y_{t-k}\right)^2}\sqrt{\sum_{k=1}^{K}\lambda_k^2\left(b_k^{(n,\tau)} - \varphi_k^{(\tau)}\right)^2},\end{aligned}$$

using the Cauchy-Schwarz inequality in the last step. Given that $\left|b^{(n,\tau)} - \varphi^{(\tau)}\right|_{\mathcal{E}} = \sqrt{\sum_{k=1}^{K}\lambda_k^2\left(b_k^{(n,\tau)} - \varphi_k^{(\tau)}\right)^2}$, this gives the result of the lemma after division by $\tau\left|b^{(n,\tau)} - \varphi^{(\tau)}\right|_{\mathcal{E}}$. ■

The next lemma establishes the relation between the constrained and penalized estimator, and together with Lemma 7 it will be used to establish a bound for the distance between the sample and population penalized estimator under the RKHS norm.

Lemma 8 *Suppose that $\varphi \in \operatorname{int}\left(\mathcal{E}\left(B\right)\right)$. Under Condition 1, if $a \in \mathcal{E}_K\left(1\right)$, and $b^{(n,\tau)}$ is as in (4), there is a $\tau = \tau_n = O_p\left(n^{-1/2}\right)$ such that $\left|b^{(n,\tau)}\right|_{\mathcal{E}} < B$ with probability going to one, and*

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(Y_t - X_t\left(b^{(n,\tau)}\right)\right)X_t\left(a\right) = O_p\left(B\sqrt{\sum_{k=1}^{K}\lambda_k^2 a_k^2}\right),$$

where the above bound holds uniformly in $a \in \mathcal{E}_K\left(1\right)$. In consequence, there is a $\tau = O_p\left(n^{-1/2}\right)$ such that $b^{(n,\tau)} = b^{(n)}$.

Moreover, for any $\tau > 0$,

$$\sqrt{\sum_{k=1}^{K} \frac{1}{\tau^2 \lambda_k^2} \left(\frac{1}{n} \sum_{t=1}^{n} \left(1 - \mathbb{E}\right) \left(Y_t - X_t \left(\varphi^{(\tau)}\right)\right) Y_{t-k} \right)^2} = O_p \left(\tau^{-1} n^{-1/2}\right).$$

Proof. Suppose that $\tau > 0$ as otherwise, by the first order conditions, the r.h.s. in the first display in the statement of lemma is exactly zero and there is nothing to prove. For arbitrary $\tau > 0$, the first order conditions that define (4) imply that

$$b_k^{(n,\tau)} = -\frac{1}{\tau \lambda_k^2} \frac{1}{n} \sum_{t=1}^{n} \left(Y_t - X_t \left(b^{(n,\tau)}\right)\right) Y_{t-k} \tag{23}$$

where $b_k^{(n,\tau)}$ is the k^{th} element in $b^{(n,\tau)}$. By Condition 1, multiplying both sides by $\tau \lambda_k^2 a_k$, with $a \in \mathcal{E}_K(1)$, and summing over k,

$$\begin{aligned} \left| \frac{1}{n} \sum_{t=1}^{n} \left(Y_t - X_t \left(b^{(n,\tau)}\right)\right) X_t(a) \right| &= \tau \left| \sum_{k=1}^{K} \lambda_k^2 b_k^{(n,\tau)} a_k \right| \\ &\leq \tau \sqrt{\sum_{k=1}^{K} \lambda_k^2 \left| b_k^{(n,\tau)} \right|^2} \sqrt{\sum_{k=1}^{K} \lambda_k^2 a_k^2}, \end{aligned} \tag{24}$$

recalling the definition of $X_t(a)$ and using the Cauchy-Schwarz inequality. Given that $a \in \mathcal{E}_K(1)$, deduce that $\sqrt{\sum_{k=1}^{K} \lambda_k^2 a_k^2} \leq 1$. Moreover, the above display clearly holds uniformly in $a \in \mathcal{E}_K(1)$. We need to show that there is a $\tau = \tau_n = O_p\left(n^{-1/2}\right)$ such $\sqrt{\sum_{k=1}^{K} \lambda_k^2 \left| b_k^{(n,\tau)} \right|^2} < B$. This will imply the first display in the statement of the lemma.

By the triangle inequality,

$$\sqrt{\sum_{k=1}^{K} \lambda_k^2 \left| b_k^{(n,\tau)} \right|^2} \leq \sqrt{\sum_{k=1}^{K} \lambda_k^2 \left| \varphi_k^{(\tau)} \right|^2} + \sqrt{\sum_{k=1}^{K} \lambda_k^2 \left(b_k^{(n,\tau)} - \varphi_k^{(\tau)} \right)^2}. \tag{25}$$

For $\tau \geq 0$, we have that $\sqrt{\sum_{k=1}^{K} \lambda_k^2 \left| \varphi_k^{(\tau)} \right|^2} \leq \sqrt{\sum_{k=1}^{K} \lambda_k^2 \left| \varphi_k \right|^2}$, because the penalized population estimator must have norm no larger than the unpenalized population estimator $\varphi^{(K)}$. But, $\varphi^{(K)}$ is restricted to be in $\mathcal{E}_K \subset \mathcal{E}$, hence its norm cannot be larger than the one of φ. By this remark and the fact that $\varphi \in \operatorname{int}\left(\mathcal{E}(B)\right)$, there is an $\epsilon > 0$

such that the first term on the r.h.s. is $B-3\epsilon$. Lemma 7 gives

$$\begin{aligned}
&\sum_{k=1}^{K}\lambda_k^2\left(b_k^{(n,\tau)}-\varphi_k^{(\tau)}\right)^2\\
\leq\ &\sum_{k=1}^{K}\frac{2}{\tau^2\lambda_k^2}\left[\frac{1}{n}\sum_{t=1}^{n}\left(Y_t-X_t\left(\varphi^{(\tau)}\right)\right)Y_{t-k}-\mathbb{E}\left(Y_t-X_t\left(\varphi^{(\tau)}\right)\right)Y_{t-k}\right]^2. \qquad (26)
\end{aligned}$$

Adding and subtracting $(1-\mathbb{E})\,X_t(\varphi)\,Y_{t-k}$, and then using the basic inequality $(x+y)^2\leq 2x^2+2y^2$ for any real x,y, the r.h.s. is

$$\begin{aligned}
&\sum_{k=1}^{K}\frac{2}{\tau^2\lambda_k^2}\left[\frac{1}{n}\sum_{t=1}^{n}(1-\mathbb{E})\left(Y_t-X_t(\varphi)\right)Y_{t-k}+\frac{1}{n}\sum_{t=1}^{n}(1-\mathbb{E})\left(X_t(\varphi)-X_t\left(\varphi^{(\tau)}\right)\right)Y_{t-k}\right]^2.\\
\leq\ &\sum_{k=1}^{K}\frac{4}{\tau^2\lambda_k^2}\left[\frac{1}{n}\sum_{t=1}^{n}(1-\mathbb{E})\left(Y_t-X_t(\varphi)\right)Y_{t-k}\right]^2\\
&+\sum_{k=1}^{K}\frac{4}{\tau^2\lambda_k^2}\left[\frac{1}{n}\sum_{t=1}^{n}(1-\mathbb{E})\left(X_t(\varphi)-X_t\left(\varphi^{(\tau)}\right)\right)Y_{t-k}\right]^2.
\end{aligned}$$

Recalling that our goal is to bound the second term on the r.h.s. of (25), the above two displays imply that

$$\begin{aligned}
\sqrt{\sum_{k=1}^{K}\lambda_k^2\left(b_k^{(n,\tau)}-\varphi_k^{(\tau)}\right)^2}\ \leq\ &\frac{1}{\tau}\sqrt{\sum_{k=1}^{K}\frac{4}{\lambda_k^2}\left[\frac{1}{n}\sum_{t=1}^{n}(1-\mathbb{E})\left(Y_t-X_t(\varphi)\right)Y_{t-k}\right]^2}\\
&+\frac{1}{\tau}\sqrt{\sum_{k=1}^{K}\frac{4}{\lambda_k^2}\left[\frac{1}{n}\sum_{t=1}^{n}(1-\mathbb{E})\left(X_t(\varphi)-X_t(\varphi^{(\tau)})\right)Y_{t-k}\right]^2}\\
=:\ &I+II. \qquad (27)
\end{aligned}$$

To bound I on the r.h.s. note that for $k>0$,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}(1-\mathbb{E})\left(Y_t-X_t(\varphi)\right)Y_{t-k}\right]^2 &= \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\varepsilon_tY_{t-k}\right]^2\\
&= \frac{\sigma^2\gamma(0)}{n}
\end{aligned}$$

(recall $\gamma(k)$ is the ACF of the AR process). Hence,

$$\sum_{k=1}^{K}\frac{1}{\lambda_k^2}\left[\frac{1}{n}\sum_{t=1}^{n}(1-\mathbb{E})(Y_t-X_t(\varphi))Y_{t-k}\right]^2=O_p\left(\frac{\sigma^2\gamma(0)}{n}\right)$$

because the coefficients λ_k^{-2} are summable. We deduce that it is possible to find a $\tau=O_p\left(n^{-1/2}\right)$ such that $I\leq\epsilon$. To bound II, recall that $\varphi^{(\tau)},\varphi\in\mathcal{E}(B)$ for any $\tau\geq 0$, and write

$$W_{k,l}:=\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(1-\mathbb{E})Y_{t-l}Y_{t-k}$$

for ease of notation. Recall that η is the exponent in Condition 1. Then, for $\rho=(2\eta+1)/2>1$,

$$\begin{aligned}III &:= \mathbb{E}\sum_{k=1}^{K}\frac{1}{\lambda_k^2}\left[\frac{1}{n}\sum_{t=1}^{n}(1-\mathbb{E})\left(X_t(\varphi)-X_t\left(\varphi^{(\tau)}\right)\right)Y_{t-k}\right]^2\\ &\leq \sum_{k=1}^{K}\frac{1}{\lambda_k^2}\mathbb{E}\sup_{b\in\mathcal{E}(2B)}\left[\frac{1}{n}\sum_{t=1}^{n}(1-\mathbb{E})\sum_{l=1}^{\infty}b_lY_{t-l}Y_{t-k}\right]^2\\ &\leq \frac{1}{n}\sum_{k=1}^{K}\frac{1}{\lambda_k^2}\sum_{l,j=1}^{\infty}l^{-\rho}j^{-\rho}\mathbb{E}W_{k,l}W_{k,j}\\ &\lesssim \frac{1}{n}\sup_{k,l,j}\mathbb{E}W_{k,l}W_{k,j}\leq\frac{1}{n}\sup_{k,l}\mathbb{E}W_{k,l}^2 \qquad (28)\end{aligned}$$

using Lemma 3 in the second inequality and summability of the coefficient in the penultimate inequality. By Lemma 4, $\mathbb{E}W_{k,l}^2\leq c$ for some finite absolute constant c. Hence, deduce that $III=O_p(n^{-1})$, which implies that $II=O_p\left(\tau^{-1}n^{-1/2}\right)$. Hence, there is a $\tau=O_p\left(n^{-1/2}\right)$ such that $II\leq\epsilon$. The control of $I+II$ implies that, with probability going to one, (27) is not greater than 2ϵ for suitable τ. Hence, we have shown that there is a $\tau=O_p\left(n^{-1/2}\right)$ such that (25) is not greater than $B-\epsilon$ with probability going to one. This bound for (25) together with (24) proves the first display in the lemma. To see that this also implies that there is a $\tau=O_p\left(n^{-1/2}\right)$ such that $b^{(n,\tau)}=b^{(n)}$ note that $\left|b^{(n,\tau)}\right|_{\mathcal{E}}$ is non-decreasing as $\tau\to 0$. Hence, $b^{(n,\tau)}=b^{(n)}$ for the smallest τ such that $\left|b^{(n,\tau)}\right|_{\mathcal{E}}\leq B$

The last statement in the lemma follows from the just derived bound for the r.h.s. of (26). ■

We now estimate the approximation error.

Lemma 9 *Suppose Condition 1. For any $K \to \infty$, we have that $\left|\varphi^{(K)} - \varphi^{(\tau)}\right|_{\mathcal{E}} \to 0$ as $\tau \to 0$ where $\varphi^{(K)}$ and $\varphi^{(\tau)}$ are as in (6) and (21). Moreover, $\left|\varphi^{(K)} - \varphi^{(\tau)}\right|_{\mathcal{E}} = O_p\left(\tau K^{2\eta}\right)$.*

Proof. The first part of the lemma is the continuity of the penalized estimator w.r.t. τ, under the RKHS norm. This is given in Theorem 5.17 in Steinwart and Christmann (2008). Hence, we only need to prove the second statement. Let Γ be the $K \times K$ matrix with (k,l) entry equal to $\gamma(k-l)$ and let Γ_1 be the first column in Γ. Let $\tilde{\varphi}^{(K)}, \tilde{\varphi}^{(\tau)} \in \mathbb{R}^K$ be the first K entries in $\varphi^{(K)}, \varphi^{(\tau)} \in \mathcal{E}_K$. Recall that in both $\varphi^{(K)}$ and $\varphi^{(\tau)}$ all entries $k > K$ are zero. Then, $\tilde{\varphi}^{(K)} = \Gamma^{-1}\Gamma_1$, and writing $D := \tau^{1/2}\Lambda$ with Λ as in (5),

$$\tilde{\varphi}^{(\tau)} = (DD + \Gamma)^{-1}\Gamma_1.$$

By the Woodbury identity (Petersen and Pedersen, 2012, eq.159)

$$(DD + \Gamma)^{-1} = \Gamma^{-1} - \Gamma^{-1}D\left(I + D\Gamma^{-1}D\right)^{-1}D\Gamma^{-1}$$

we have that

$$\tilde{\varphi}^{(K)} - \tilde{\varphi}^{(\tau)} = \left[\Gamma^{-1}D\left(I + D\Gamma^{-1}D\right)^{-1}D\Gamma^{-1}\right]\Gamma_1.$$

We also have that $\left|\varphi^{(K)} - \varphi^{(\tau)}\right|_{\mathcal{E}} = \left|\Lambda\left(\tilde{\varphi}^{(K)} - \tilde{\varphi}^{(\tau)}\right)\right|_2$. To keep the notation simple, we are using $|\cdot|_2$ for the ℓ_2 norm as well as for the Euclidean norm in $\mathbb{R}^K$. Hence, using the above display,

$$\begin{aligned}\left|\varphi^{(K)} - \varphi^{(\tau)}\right|_{\mathcal{E}} &= \left|\Lambda\Gamma^{-1}D\left(I + D\Gamma^{-1}D\right)^{-1}D\Gamma^{-1}\Gamma_1\right|_2 \\ &= \left|D\Gamma^{-1}D\left(I + D\Gamma^{-1}D\right)^{-1}\Lambda\tilde{\varphi}^{(K)}\right|_2\end{aligned}$$

using the definitions of $\tilde{\varphi}^{(K)}$ and D in the second equality. For any square matrix W and a compatible vector a, $|Wa|_2 \leq \sigma_{\max}(W)\,|a|_2$, where $\sigma_{\max}(W)$ is the maximum singular value of W. There is a $B < \infty$ such that $\varphi \in \mathcal{E}(B)$. Because $\left|\varphi^{(K)}\right|_{\mathcal{E}} \leq |\varphi|_{\mathcal{E}} \leq B$, we then must have that $|\Lambda\tilde{\varphi}|_2 \leq B$. Hence, we only need to find the maximum singular value of $W = D\Gamma^{-1}D\left(I + D\Gamma^{-1}D\right)^{-1}$. Using the basic equality $AC^{-1} = \left(CA^{-1}\right)^{-1}$ for invertible $A = D\Gamma^{-1}D$ and $C = \left(I + D\Gamma^{-1}D\right)$, we have that $W = \left(D^{-1}\Gamma D^{-1} + I\right)^{-1}$. The matrix $\left(I + D^{-1}\Gamma D^{-1}\right)$ has eigenvalues equal to 1 plus the eigenvalues of $D^{-1}\Gamma D^{-1}$. Hence, we focus on finding the smallest singular value of $D^{-1}\Gamma D^{-1}$. The following inequalities hold for the singular values of the product of two matrices A and C:

$$\sigma_{\min}(A)\,\sigma_{\min}(C) \leq \sigma_{\min}(AC) \leq \sigma_{\max}(AC) \leq \sigma_{\max}(A)\,\sigma_{\max}(C)$$

where $\sigma_{\max}(\cdot)$ and $\sigma_{\min}(\cdot)$ are the maximum and minimum singular value of the matrix argument (Bathia, 1997, eq. III.20, p.72). Recall that, in order to derive (20), we argued that Γ has minimum eigenvalue $\theta_{\min}$ bounded away from zero. Moreover, D^{-1} has minimum eigenvalues equal to a constant multiple of $\tau^{-1/2}K^{-\eta}$. Hence, applying the inequalities above with $A = D^{-1}$ and $C = \Gamma D^{-1}$ we have that $D^{-1}\Gamma D^{-1}$ has eigenvalues bounded below by a constant multiple of $\tau^{-1}\theta_{\min}K^{-2\eta}$. This implies that $\sigma_{\max}(W) \lesssim \left(1+\theta_{\min}\tau^{-1}K^{-2\eta}\right)^{-1}$. Hence, after rearrangement, deduce that $\left|\varphi^{(K)} - \varphi^{(\tau)}\right|_{\mathcal{E}} \lesssim \tau K^{2\eta}\left(\theta_{\min} + \tau K^{2\eta}\right)^{-1}$. This is $O\left(\tau K^{2\eta}\right)$ as stated in the lemma. ■

We need a final approximation result.

Lemma 10 *Recall (6). Suppose Condition 1. If $\varphi \in \mathcal{E}$, then $\left|\varphi^{(K)} - \varphi\right|_{\mathcal{E}} = o(1)$ as $K \to \infty$. If also $|\varphi_k| \lesssim k^{-\nu}$ with $\nu > (2\eta+1)/2$, then, $\left|\varphi^{(K)} - \varphi\right|_{\mathcal{E}} = O\left(K^{(2\eta+1-2\nu)/2}\right)$.*

Proof. Recall the definition of $\beta = \beta^{(K)} \in \mathbb{R}^{\infty}$ just before (17). Let $\tilde{\beta} \in \mathbb{R}^K$ have the same first K entries as β. Write $Y_t = X_t(\beta) + \varepsilon_{K,t}$ where $\varepsilon_{K,t} = \varepsilon_t - X_t(\beta - \varphi)$. Given that $\tilde{\varphi}^{(K)}$ is the population ordinary least square estimator, using the same notation as in the proof of Lemma 9,

$$\tilde{\varphi}^{(K)} = \tilde{\beta} + \Gamma^{-1}\mathbb{E}\begin{pmatrix} Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-K} \end{pmatrix}\varepsilon_{K,t}.$$

We need to show that the second term goes to zero under the norm $|\cdot|_{\mathcal{E}}$. Given that the innovations ε_t are i.i.d., the expectation is equal to

$$-\mathbb{E}\begin{pmatrix} Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-K} \end{pmatrix}\sum_{l=K+1}^{\infty} Y_{t-l}\varphi_l = -\sum_{l=1}^{\infty}\varphi_{K+l}\begin{pmatrix} \gamma(K-1+l) \\ \gamma(K-2+l) \\ \vdots \\ \gamma(l) \end{pmatrix} =: \Psi.$$

Hence,

$$\left|\beta - \varphi^{(K)}\right|_{\mathcal{E}} = \left|\Lambda\Gamma^{-1}\Psi\right|_2,$$

and again we are using $|\cdot|_2$ for the ℓ_2 norm as well as for the Euclidean norm in $\mathbb{R}^K$. We need to show that this converges to zero. By similar arguments as in the proof of Lemma 9, $|\Lambda\Gamma^{-1}\Psi|_2 \lesssim K^{\eta}\theta_{\min}^{-1}|\Psi|_2$ because Λ is diagonal with largest entry $O(K^{\eta})$. To bound $|\Psi|_2$, note that

$$|\Psi|_2^2 = \sum_{l_1,l_2=1}^{\infty} \varphi_{K+l_1}\varphi_{K+l_2}\sum_{k=1}^{K}\gamma(K-k+l_1)\gamma(K-k+l_2).$$

However, $\max_{k\leq K}|\gamma(K-k+l)| \lesssim |\gamma(l)|$ because the ACF is absolutely summable by Lemma 1. Hence, $|\Psi|_2^2 \lesssim K\varphi_K^2$ and from the previous remarks, deduce that

$$\left|\beta-\varphi^{(K)}\right|_{\mathcal{E}} \lesssim \sqrt{K^{2\eta+1}\varphi_K^2}.$$

Because $\varphi \in \mathcal{E}$, then $|\varphi_K|_{\mathcal{E}} = O\left(K^{-(2\eta+1)/2}/\ln^{1/2}(1+K)\right)$ by Lemma 2, so that $\left|\beta-\varphi^{(K)}\right|_{\mathcal{E}} = o(1)$. If also $|\varphi_k| \lesssim k^{-\nu}$ holds true, $\left|\varphi^{(K)}-\beta\right|_{\mathcal{E}} \lesssim K^{(2\eta+1-2\nu)/2}$. By definition of β,

$$|\varphi-\beta|_{\mathcal{E}} = \sqrt{\sum_{k>K}\varphi_k^2\lambda_k^2} = \begin{cases} O\left(K^{(2\eta+1-2\nu)/2}\right) & \text{if } |\varphi_k| \lesssim k^{-\nu} \\ o(1) & \text{if } \varphi \in\mathcal{E} \end{cases}.$$

Hence, by the triangle inequality, we deduce the bound for $\left|\varphi^{(K)}-\varphi\right|_{\mathcal{E}}$. ■

Proof of Theorem 1 Points 2-6. We can now prove Points 2-6 in Theorem 1.

Point 2. If $\varphi \in \mathcal{E}$, then, there is a finite B such that $\varphi \in \operatorname{int}(\mathcal{E}(B))$. By Lemma 7 and 8, deduce that $\left|b^{(n,\tau)}-\varphi^{(\tau)}\right|_{\mathcal{E}} = O_p\left(\tau^{-1}n^{-1/2}\right)$. Hence, if $\tau n^{1/2}\to\infty$ in probability, by Lemma 9 and the triangle inequality, $\left|b^{(n,\tau)}-\varphi^{(K)}\right|_{\mathcal{E}} \to 0$ in probability for any $K>0$, including $K\to\infty$. By Lemma 10, $\left|\varphi-\varphi^{(K)}\right|_{\mathcal{E}} \to 0$ as long as $K\to\infty$. In consequence, the triangle inequality implies that $\left|b^{(n,\tau)}-\varphi\right|_{\mathcal{E}} \to 0$ in probability, under the sole condition that $\tau n^{1/2}+K\to\infty$ in probability.

Point 3. This follows from Lemma 8.

Point 4. By the triangle inequality, $\left|b^{(n,\tau)}-\varphi^{(K)}\right|_{\mathcal{E}} \leq \left|b^{(n,\tau)}-\varphi^{(\tau)}\right|_{\mathcal{E}}+\left|\varphi^{(K)}-\varphi^{(\tau)}\right|_{\mathcal{E}}$. Use Lemma 9 to bound the second term on the r.h.s. by a quantity $O_p\left(\tau K^{2\eta}\right)$. Use Lemmas 7 and 8 to bound the first term on the r.h.s. by a quantity $O_p\left(\tau^{-1}n^{-1/2}\right)$. Deduce that $\left|b^{(n,\tau)}-\varphi^{(K)}\right|_{\mathcal{E}} = O_p\left(\tau^{-1}n^{-1/2}+\tau K^{2\eta}\right)$. Equating the two terms inside $O_p(\cdot)$, and solving for τ, this quantity is $O_p\left(n^{-1/4}K^{\eta}\right)$ when $\tau \asymp n^{-1/4}K^{-\eta}$.

Point 5. The approximation rates are from Lemma 10.

Point 6. Lemma 8 shows that for the constrained problem, the Lagrange multiplier is $\tau = \tau_{n,B} = O_p\left(n^{-1/2}\right)$, and the constraint is possibly binding. In fact, there is a K large enough relatively to n, such that the constraint needs to be binding so that $\left|b^{(n)}\right|_{\mathcal{E}} = B$. However, if $\varphi \in \operatorname{int}\left(\mathcal{E}\left(B\right)\right)$ there is an $\epsilon > 0$ such that $\left|\varphi\right|_{\mathcal{E}} = B - \epsilon$. Then, we must have that

$$\begin{aligned}\left|b^{(n)} - \varphi\right|_{\mathcal{E}}^2 &= \left|b^{(n)}\right|_{\mathcal{E}}^2 + \left|\varphi\right|_{\mathcal{E}}^2 - 2\left\langle b^{(n)}, \varphi\right\rangle_{\mathcal{E}} \\ &= \left(B^2 + \left(B-\epsilon\right)^2 - 2\left\langle b^{(n)}, \varphi\right\rangle_{\mathcal{E}}\right).\end{aligned}$$

But $\left\langle b^{(n)}, \varphi\right\rangle_{\mathcal{E}} \leq \left|b^{(n)}\right|_{\mathcal{E}} \left|\varphi\right|_{\mathcal{E}} \leq B\left(B-\epsilon\right)$. Therefore, the above display is greater or equal than

$$B^2 + \left(B-\epsilon\right)^2 - 2B\left(B-\epsilon\right) \geq \epsilon^2.$$

This means that $b^{(n)}$ cannot converge under the norm $\left|\cdot\right|_{\mathcal{E}}$.

5.2 Proof of Corollary 1

By Points 4-5 in Theorem 1 and the triangle inequality, deduce that $\left|b^{(n,\tau)} - \varphi\right|_{\mathcal{E}} = O_p\left(n^{-1/4}K^{\eta} + K^{(2\eta+1-2\nu)/2}\right)$. Equating the coefficients this is $O_p\left(n^{-\frac{2\nu-(2\eta+1)}{4(2\nu-1)}}\right)$ when $K = n^{\frac{1}{2(2\nu-1)}}$.

References

[1] Bathia, R. (1997) Matrix Analysis. New York: Springer.

[2] Brockwell, P.J. and A. Davis (1991) Time Series: Theory and Methods. New York: Springer.

[3] Bühlmann, P. (1995). Moving-Average Representation for Autoregressive Approximations. Stochastic Processes and their Applications 60, 331-342.

[4] Bühlmann, P. (1997) Sieve Bootstrap for Time Series. Bernoulli 3, 123-148.

[5] Burman, P. and D. Nolan (1992) Data-Dependent Estimation of Prediction Functions. Journal of Time Series Analysis 13, 189-207.

[6] Burman, P., E. Chow and D. Nolan (1994) A Cross-Validatory Method for Dependent Data. Biometrika 81, 351-358.

[7] Fu, W. and K. Knight (2000) Asymptotics for Lasso-Type Estimators. Annals of Statistics 28, 1356-1378.

[8] Geyer, C.J. (1994) On the Asymptotic of Constrained M-Estimation. Annals of Statistics 22, 1993-2010.

[9] Graf, S. and H. Luschgy (2004) Sharp Asymptotics of the Metric Entropy for Ellipsoids. Journal of Complexity 20, 876-882.

[10] Györfi, L., W. Härdle, P. Sarda and P. Vieu (1990) Nonparametric Curve Estimation from Time Series. Heidelberg: Springer.

[11] Györfi, L. and A. Sancetta (2015) An open problem on strongly consistent learning of the best prediction for Gaussian processes. in M. Akritas, S.N. Lahiri and D. Politis (eds.), Proceedings of the first conference of the international Society of Nonparametric Statistics. Heidelberg: Springer.

[12] Hastie, T, R. Tibshirani and J. Friedman (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.

[13] Kreiss, J.-P., E. Paparoditis, and D.N. Politis (2011) On the Range of Validity of the Autoregressive Sieve Bootstrap. Annals of Statistics 39, 2103-2130.

[14] Kuelbs, J. (1976) A Strong Convergence Theorem for Banach Space Valued Random Variables. Annals of Probability 4, 744-771.

[15] Li, W. and W. Linde (1999) Approximation, Metric Entropy and Small Ball Estimates for Gaussian Measures. Annals of Probability 27 1556-1578.

[16] Lutz, L.W. and P. Bühlmann (2006) Boosting for High-Multivariate Responses in High Dimensional Linear Regression. Statistica Sinica 16, 471-494.

[17] Petersen, B. and M.S. Pedersen (2012) The Matrix Cookbook. URL:www.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf

[18] Schäfer, D. (2002) Strongly Consistent Online Forecasting of Centered Gaussian Processes. IEEE Transactions on Information Theory 48, 791-799.

[19] Sims, C. (1980) Macroeconomics and Reality. Econometrica 48, 1-48.

[20] Steinwart, I., and A. Christmann (2008) Support Vector Machines. Berlin: Springer.

[21] van der Vaart, A. and J.A. Wellner (2000) Weak Convergence and Empirical Processes. New York: Springer.