Electronic Journal of Statistics
Vol. 9 (2015) 801–810
ISSN: 1935-7524
DOI: 10.1214/15-EJS1007

Discussion of "High-dimensional autocovariance matrices and optimal linear prediction"*,†

Xiaohui Chen

University of Illinois at Urbana-Champaign
e-mail: xhchen@illinois.edu

Keywords and phrases: Optimal linear predition, shrinkage, sparsity.

Received February 2015.

First, we would like to congratulate Prof. McMurry and Prof. Politis for their thought-provoking paper on the optimal linear prediction based on full time series sample (hereafter, referred as [MP15]). [MP15] considered the one-step optimal linear predictor $X_{n+1}^* = \sum_{i=1}^n \phi_i(n) X_{n+1-i}$ of a univariate time series $X_1, \ldots, X_n$ in the ℓ^2 sense which is given by the solution of the Yule-Walker equations

$$\phi(n) = \Gamma_n^{-1}\gamma(n).$$

[MP15] constructed an optimal linear predictor using the full sample path

$$\hat{\phi}(n) = \hat{\Gamma}_n^{-1}\hat{\gamma}(n),$$

where $\hat{\Gamma}_n$ is a flat-top tapered sample autocovariance matrix and $\hat{\gamma}(n)$ is the shifted first row of $\hat{\Gamma}_n$. Under mild assumptions, it is shown that $\hat{\phi}(n)$ is an ℓ^2 consistent estimator of $\phi(n)$ and the resulting optimal linear prediction is consistent for X_{n+1}^*. Since computing $\hat{\phi}(n)$ requires the inversion of $\hat{\Gamma}_n$, which can be either non-positive-definite due to the matrix tapering or numerically ill-conditioned in view of the large dimension n of the full sample autocovariance matrix, [MP15] proposed four positive definiteness (pd) corrections by thresholding on bad eigenvalues or shrinking towards certain positive-definite targets. Below, we propose an alternative simple correction method that does not require the pd correction by directly working on the shrinkage of the inverse.

Let $\hat{\Gamma}_n = Q\hat{D}Q^\top$ be the eigen-decomposition of the flat-top tapered sample autocovariance matrix of $X_1, \ldots, X_n$ and $\hat{D} = \mathrm{diag}(\hat{d}_1, \ldots, \hat{d}_n)$ be the diagonal matrix containing the real eigenvalues of $\hat{\Gamma}_n$. Let

$$\hat{\Theta}_n = Q\hat{H}Q^\top,$$

*Main article 10.1214/15-EJS1000.

†Research supported by NSF grant DMS-1404891 and UIUC Research Board Award RB15004.

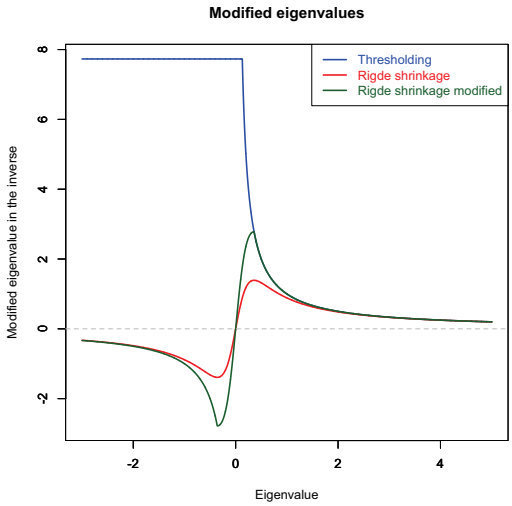

FIG 1. *Modified eigenvalues of the thresholded positive definite version of* $\hat{\Gamma}_n$ *in Section 4.1 of McMurry & Politis (2015) [MP15] and the ridge shrinkage matrix* $\hat{\Theta}_n$. *Plot from AR(1) model with coefficient* -0.5, *threshold parameter* $\lambda = d_i^{\epsilon} = \epsilon\hat{\gamma}_0/n^{\beta}$ *with* $\epsilon = 20$ *and* $\beta = 1$.

where $\hat{H} = \text{diag}(\hat{d}_1/(\hat{d}_1^2+\lambda), \ldots, \hat{d}_n/(\hat{d}_n^2+\lambda))$ and λ is a nonnegative shrinkage tuning parameter. Then, our optimal linear predictor coefficient $\phi(n)$ is defined as

$$\tilde{\phi}(n) = \hat{\Theta}_n \hat{\gamma}(n).$$

It is noted that $\tilde{\phi}(n)$ can be viewed as a ridge-type shrinkage modification of $\hat{\Gamma}_n^{-1}$ since by the orthogonality of Q we may write $\hat{\Theta}_n = (\hat{\Gamma}_n^{\top}\hat{\Gamma}_n + \lambda I_n)^{-1}\hat{\Gamma}_n^{\top}$. In particular, for fixed λ, if $\hat{d}_i \to 0$, then $\hat{h}_i = \hat{d}_i/(\hat{d}_i^2+\lambda) \sim \hat{d}_i/\lambda$; if $|\hat{d}_i| \to \infty$, then $\hat{h}_i \sim \hat{d}_i^{-1}$. Different from the thresholding and pd shrinkage corrections proposed by [MP15], our proposal $\tilde{\phi}(n)$ allows $\hat{d}_i \leq 0$ after the eigenvalue correction. Therefore, our correction achieves comparable performance for large eigenvalues of $\hat{\Sigma}_n$ as the thresholding (or the pd correction), but it can be numerically more stable when eigenvalues are below a threshold. The effect on the modified eigenvalues of $\hat{\Gamma}_n^{-1}$ is shown in Figure 1. We also include a variant of the ridge shrinkage correction on the eigenvalues

$$\tilde{h}_i = \begin{cases} \hat{d}_i^{-1} & \text{if } |\hat{d}_i| \geq \lambda^{1/2} \\ 2\hat{d}_i/(\hat{d}_i^2+\lambda) & \text{if } |\hat{d}_i| < \lambda^{1/2} \end{cases}.$$

In the following theorem, we show that the ridge corrected estimator $\tilde{\phi}(n)$, as well as its modified version using $\hat{\Theta}_n = Q\tilde{H}Q^{\top}$, both achieve the same rate of convergence for estimating $\phi(n)$ as in [MP15]. For the modified version of ridge shrinkage, since with high probability $\tilde{H} = D^{-1}$ and $\hat{\Theta}_n^{-1}$ maintains the banded structure in $\hat{\Gamma}_n$, therefore it also has the predictive consistency.

Theorem 1. *Under Assumptions 1–4 of [MP15] and if* $\lambda = o(1)$*, then we have* $|\tilde{\phi}(n) - \phi(n)|_2 = O_p(r_n + \lambda)$*, where* $r_n = ln^{-1/2} + \sum_{i=l}^{\infty} |\gamma_i|$*. In addition, under Assumptions 1–6 of [MP15] and if* $\lambda = O(r_n)$*, then for the modified version of the ridge correction* $\hat{\Theta}_n = Q\tilde{H}Q^\top$*, we have* $|\tilde{X}_{n+1} - X^*_{n+1}| = o(1)$*.*

For a practical choice of λ, we can use $\lambda = \epsilon\hat{\gamma}_0/n^\beta$ where $\epsilon > 0$ and $\beta > 1/2$ are positive constants. Then, following a similar argument in [MP10], we can easily show that the same asymptotic rate of convergence and prediction consistency in Theorem 1 are attained for such choice. Note that the value of λ has the same form as the thresholding parameter in [MP15]. In our simulation examples, we set $\epsilon = 20$ and $\beta = 1$.

1. Sparse full-sample optimal linear prediction

For optimal linear predictors that are high order yet may be approximated by a only few large coefficients in $\phi(n)$, it is expected that the sparse approximation may work better than using the full sample path. For example, consider the optimal linear predictor $X_i = \sum_{j=1}^{14} \phi_j X_{i-j} + e_i$, where $\phi_j = 0$ except that $\phi_1 = -0.3, \phi_3 = 0.7, \phi_{14} = -0.2$. This observation leads to our second proposal of a sparse full-sample optimal (SFSO) linear predictor $\hat{\phi}^{SFSO}$

$$\begin{aligned} \text{minimize}_{b\in\mathbb{R}^n} \quad & |b|_1 \\ \text{subject to} \quad & |\hat{\Gamma}_n b - \hat{\gamma}(n)|_\infty \le \lambda. \end{aligned}$$

The SFSO predictor is a Dantzig-selector type method in high-dimensional linear models and estimation of sparse precision matrix [CT07, CLL11]. The SFSO is computationally efficient since it can be recast to a linear program (LP)

$$\begin{aligned} \text{minimize}_{b^+,b^-\in\mathbb{R}^n} \quad & \sum_{i=1}^{n} b_i^+ + b_i^- \\ \text{subject to} \quad & b^+, b^- \ge 0 \\ & \hat{\Gamma}_n b^+ - \hat{\Gamma}_n b^- \le \lambda 1_n + \hat{\gamma}(n) \\ & -\hat{\Gamma}_n b^+ + \hat{\Gamma}_n b^- \le \lambda 1_n - \hat{\gamma}(n), \end{aligned}$$

where $1_n = (1, \ldots, 1)^\top$ of size n. Let $(\hat{b}^+, \hat{b}^-)$ be the solution of the above LP. Then $\hat{\phi}^{SFSO} = \hat{b}^+ - \hat{b}^-$. Due to the sparsity-promoting nature of the ℓ^1 norm, the SFSO can simultaneously perform predictor selection, estimation and one-step prediction. Statistically, there is another important advantage of SFSO over AR models with order determined by the AIC. In AR models, once the order is determined, predictors have to be added into the model in the sequential way, therefore necessitating a large model size if the sparse non-zeros are not ordered in the time index. In the above example of AR(14) with three non-zero coefficients, even with a correctly specified AR model, we need a model with 14 predictors in order to achieve the optimal prediction! In contrast, the SFSO does not depend on the order of predictors and therefore it has more flexibility

of selecting the predictors. Let $r \in [0,1)$. We consider the following class of sparse vectors

$$\mathcal{G}_r(C_0, M) = \left\{ \mathbf{b} \in \mathbb{R}^n : \max_{1\le i\le n} |b_i| \le C_0, \sum_{i=1}^{n} |b_i|^r \le M \right\},$$

where M is the sparsity parameter and it is allowed to grow with the dimension n.

Theorem 2. *Let $q > 2$ and $d > 1/2 - 1/q$. Assume that $\mathbb{E}|X_t^{2q}| \le \nu < \infty$ and the functional dependence measure $\delta_{2q,t} = (\mathbb{E}|X_t - X_t'|^{2q})^{1/2q} \le C_q t^{-d-1}$. Suppose $\phi(n) \in \mathcal{G}_r(C_0, M)$. Let*

$$\lambda \ge (1+2C_0^{1-r}M) \left\{ \max\left(\frac{l^{1/q}}{n^{1-1/q}}, \sqrt{\frac{\log l}{n}} \right) + n^{-1} \max_{1\le s\le \lfloor c_\kappa l \rfloor} s|\gamma_s| + \max_{l<s\le n} |\gamma_s| \right\}.$$

Under Assumptions 2 and 4 of [MP15], we have

$$|\hat{\phi}^{SFSO} - \phi(n)|_2 = O_P(M^{1/2} |\Gamma_n^{-1}|_{L^\infty}^{1-r/2} \lambda^{1-r/2}), \tag{1}$$

where $|A|_{L^\infty} = \max_i \sum_j |A_{ij}|$ is the matrix L^∞ norm.

Remark 1. If $\gamma_s = O(s^{-d-1})$, then $n^{-1}\max_{1\le s\le \lfloor c_\kappa l \rfloor} s|\gamma_s| = O(n^{-1})$, which is dominated by the first term in the bracket. Consider $r = 0$; then ϕ is an M-sparse vector. Comparing the rate of convergence (1) with Theorem 1 where the rate is $O(r_n), r_n = ln^{-1/2} + \sum_{i=l}^{\infty} |\gamma_i|$, we observe that better rate is obtained for the SFSO if M is constant (or slowly grows with n at proper rate) since

$$\max\left(\frac{l^{1/q}}{n^{1-1/q}}, \sqrt{\frac{\log l}{n}} \right) \ll \frac{l}{n^{1/2}} \quad \text{and} \quad \max_{l<s\le n} |\gamma_s| \ll \sum_{i=l}^{\infty} |\gamma_i|, \quad l \to \infty.$$

We can also obtain the result for all short-range dependence time series $d \in (0, 1/2 - 1/q]$. In addition, the estimation error can be obtained under the ℓ^w loss functions for all $w \in [1, \infty]$. Details are omitted.

2. Simulation examples

We now compare the finite sample performance of the proposed ridge corrected shrinkage and SFSO linear predictors with thresholding, shrinkage to a positive definite matrix and white noise proposed in [MP15]. We also run the R function `ar()` with the default parameter that uses the Yule-Walker solution with order selection by the AIC. Partially following the setups in [MP15], we consider the following three models

1. AR(1) model: $X_i = \theta X_{i-1} + e_i$, where $\theta = -0.1, -0.5, -0.9$ and e_i are iid $N(0,1)$.
2. MA(1) model: $X_i = e_i + \theta e_{i-1}$, where $\theta = -0.1, -0.5, -0.9$ and e_i are iid $N(0,1)$.

TABLE 1
RMSPE and mean ℓ^1 estimation error for the AR(14) model. FSO-Ridge and FSO-Ridge-Thr are the ridge shrinkage corrected and its modified variant in this paper. SFSO is the sparse full-sample optimal linear predictor and the rest of the symbols are consistent with [MP15]

	RMSPE	Mean ℓ^1 estimation error
AR	1.0347	0.8329
FSO-Th-Raw	1.3093	6.7437
FSO-Th-Shr	1.1987	5.2161
FSO-PD-Raw	1.2043	6.2712
FSO-PD-Shr	1.1941	7.0582
FSO-WN-Raw	1.1998	5.1059
FSO-WN-Shr	1.1986	4.2637
FSO-Ridge	1.1268	3.4535
FSO-Ridge-Thr	1.1547	4.5370
SFSO	1.0325	0.4381

3. Higher-order AR(14) model: $X_i = \sum_{j=1}^{14} \theta_j X_{i-j} + e_i$, where $\theta_1 = -0.3$, $\theta_3 = 0.7, \theta_{14} = -0.2$, and the rest of $\theta_j = 0$. The errors e_i are iid $N(0,1)$.

For AR(1), it is expected that `ar()` does the best. For MA(1), it is expected that the shrinkage type estimators would work better than the AR and SFSO linear predictors since the latter two are misspecified. For the higher-order AR(14), it is expected that the SFSO performs among the best because of the sparsity structure. The sample size is fixed to 200 for all simulations and the 201-st observation is used to test for prediction. We follow the empirical rule for choosing the bandwidth parameter l in [MP15]. The performance of those estimators are assessed by the root mean square prediction error (RMSPE) and the mean ℓ^1 estimation error. All numbers in Table 1–3 are reported by averaging 1000 simulation times. From Table 1, it is observed that the AR and SFSO predictors are comparably the best in terms of the RMSPE among all predictors considered here, followed up the FSO-ridge. The superior predictive performance of AR is conjectured due to the correct model specification. Interestingly, if we look at the estimation errors, there is a sizable improvement for the SFSO over the AR due to sparsity.

From Table 2, AR has top performances among all simulations with $\theta = -0.9, -0.5$ since it is the right model where the data are simulated. However, it is again interesting to observe that the SFSO also provides a satisfactory output (often the best) for the prediction and predictor selection. For $\theta = -0.1$, the advantage of AR becomes small since the time series are weakly dependent. From Table 3, it is not surprising to observe that FSO with proper thresholding/shrinkage can achieve better performance, though it seems that the best correction is setup-dependent. Overall, the FSO-ridge predictor has comparable performance with other FSO thresholding/shrinkage predictors in all simulation setups. We remark here that our Table 2 and 3 are slightly different from Table 1 and 2 in [MP15]. We do not know that whether or not the differences are artifacts of stochastic errors in the simulation or due to different implementations.

TABLE 2
RMSPE (columns 2–4) and mean ℓ^1 estimation error (columns 5–7) for the AR(1) models

	$\theta = -0.9$	$\theta = -0.5$	$\theta = -0.1$	$\theta = -0.9$	$\theta = -0.5$	$\theta = -0.1$
AR	1.0216	1.0371	0.9758	0.1466	0.1502	0.1514
FSO-Th-Raw	1.2024	1.0655	0.9753	2.2968	0.7728	0.1143
FSO-Th-Shr	1.1004	1.0628	0.9753	1.8906	0.7569	0.1143
FSO-PD-Raw	1.1562	1.0413	0.9753	3.1944	0.4203	0.1114
FSO-PD-Shr	1.0966	1.0371	0.9746	4.0274	0.4538	0.1066
FSO-WN-Raw	1.1557	1.0527	0.9753	1.8588	0.6363	0.1143
FSO-WN-Shr	1.1449	1.0503	0.9753	1.6644	0.6115	0.1143
FSO-Ridge	1.1085	1.0380	0.9749	1.6139	0.3548	0.1086
FSO-Ridge-Thr	1.0858	1.0488	0.9752	1.9124	0.5546	0.1143
SFSO	1.0269	1.0371	0.9738	0.1206	0.1513	0.0966

TABLE 3
RMSPE (columns 2–4) and mean ℓ^1 estimation error (columns 5–7) for the MA(1) models

	$\theta = -0.9$	$\theta = -0.5$	$\theta = -0.1$	$\theta = -0.9$	$\theta = -0.5$	$\theta = -0.1$
AR	1.0083	1.0072	0.9962	5.1003	0.3934	0.1586
FSO-Th-Raw	1.0244	1.0059	0.9938	6.9455	0.4851	0.1235
FSO-Th-Shr	1.0202	1.0052	0.9938	7.0753	0.4769	0.1235
FSO-PD-Raw	1.0502	0.9972	0.9940	7.9566	0.3735	0.1207
FSO-PD-Shr	1.0749	1.0006	0.9937	8.2909	0.5524	0.1162
FSO-WN-Raw	1.0146	1.0014	0.9938	7.2256	0.3693	0.1235
FSO-WN-Shr	1.0171	0.9999	0.9938	7.4006	0.3512	0.1235
FSO-Ridge	1.0646	0.9943	0.9936	8.4464	0.4842	0.1179
FSO-Ridge-Thr	1.0438	0.9939	0.9937	8.1716	0.3721	0.1231
SFSO	1.0854	1.0150	0.9930	8.4294	0.7175	0.1080

3. Concluding remarks

We thank Prof. McMurry and Prof. Politis for their stimulating paper which for the first time shows the feasibility of consistent optimal linear prediction based on the full sample. Motivated from their work, we proposed an alternative correction of the optimal linear prediction which has the ridge shrinkage interpretation and does not require the positive-definiteness as in [MP15]. We also proposed a sparse optimal linear predictor using the full sample (SFSO) that simultaneously performs predictor selection and one-step prediction. Asymptotic rate of convergence was established for both methods under analogous assumptions in [MP15]. In addition, prediction consistency is established for a modified version of our ridge correction method. Finite sample performances were studied by three simulation examples. We noted that the numeric performances in those examples depend on the tuning parameter λ, which was fixed in all simulations. We simply used $\lambda = 20\hat{\gamma}_0/n$ and $\sqrt{\log(n)/n}$ for ridge corrected FSO and SFSO predictors respectively. Better performance can be achieved if we tune those parameters. Tuning parameter selection is an open question for optimal prediction, as well as estimation, in high-dimensional time series analysis. Though the superior predictive performance of the SFSO linear predictor is demonstrated in the simulation under sparse settings, it is an interesting question that to what extend the SFSO has the prediction consistency. We leave this as future work.

Appendix: Proofs

Proof of Theorem 1. Note that

$$\tilde{\phi}(n) - \phi(n) = \hat{\Theta}_n(\hat{\gamma}(n) - \gamma(n)) + (\hat{\Theta}_n - \Gamma_n^{-1})\gamma(n).$$

Therefore

$$|\tilde{\phi}(n) - \phi(n)|_2 \leq \rho(\hat{\Theta}_n)|\hat{\gamma}(n) - \gamma(n)|_2 + \rho(\hat{\Theta}_n - \Gamma_n^{-1})|\hat{\gamma}(n)|_2$$

By Lemma 1 and Theorem 1 [MP15], $|\hat{\gamma}(n)-\gamma(n)|_2 = O_P(r_n)$ and $\rho(\hat{\Gamma}_n-\Gamma_n) = O_P(r_n)$. Since the spectral density of X_i is bounded between $[c_1, c_2]$, we have that all eigenvalues $\rho_i(\Gamma_n) \in [2\pi c_1, 2\pi c_2], i = 1, \ldots, n$. Let $G = \{\pi c_1 \leq \hat{d}_i \leq \pi c_2, \forall i = 1, \ldots, n\}$, where $\hat{d}_i = \rho_i(\hat{\Gamma}_n)$. Then, $\mathbb{P}(G) \to 1$ as $n \to \infty$. Since $\hat{\Gamma}_n$ is positive definite on the event G, we have with probability tending to one

$$\begin{aligned}
\rho(\hat{\Theta}_n - \Gamma_n^{-1}) &\leq \rho(\hat{\Theta}_n - \hat{\Gamma}_n^{-1}) + \rho(\hat{\Gamma}_n^{-1} - \Gamma_n^{-1}) \\
&\leq \max_{1\leq i\leq n} |\hat{h}_i - \hat{d}_i^{-1}| + O_p(r_n) \\
&= \max_{1\leq i\leq n} \frac{\lambda}{|\hat{d}_i(\hat{d}_i^2 + \lambda)|} + O_p(r_n) \\
&= O_p(\lambda + r_n).
\end{aligned}$$

By the short-range dependence assumption $\sum_{i=1}^{\infty} |\gamma_i| < \infty$, it follows that $|\tilde{\phi}(n) - \phi(n)|_2 = O_p(\lambda + r_n)$. The same rate is obtained by observing that

$$\begin{aligned}
\max_{1\leq i\leq n} |\tilde{h}_i - \hat{d}_i^{-1}| &= \max_{1\leq i\leq n} \left| \frac{2\hat{d}_i}{\hat{d}_i^2 + \lambda} - \frac{1}{\hat{d}_i} \right| 1(|\hat{d}_i| < \lambda^{1/2}) \\
&= \max_{1\leq i\leq n} \left| \frac{\lambda - \hat{d}_i^2}{\hat{d}_i(\hat{d}_i^2 + \lambda)} \right| 1(|\hat{d}_i| < \lambda^{1/2}) \\
&= O_P(\lambda).
\end{aligned}$$

Since $Q\tilde{H}^{-1}Q^\top$ has the same banded structure as $\hat{\Gamma}_n$ on G for sufficiently large n, the rest of the proof for prediction consistency follows from the argument of [MP15]. □

Proof of Theorem 2. For notation simplicity, we write $\hat{\phi} = \hat{\phi}^{SFSO}$ and $\phi = \phi(n)$. Let $G_1 = \{|\hat{\gamma}(n) - \gamma(n)|_\infty \leq \epsilon\}$ and $G_2 = \{|\hat{\Gamma}_n - \Gamma_n|_\infty \leq \epsilon\}$. Since $\gamma(n) = \Gamma_n\phi$, we have on the event $G = G_1 \cap G_2$

$$\begin{aligned}
|\hat{\Gamma}_n\phi - \hat{\gamma}(n)|_\infty &= |\hat{\Gamma}_n\phi - \gamma(n) + \gamma(n) - \hat{\gamma}(n)|_\infty \\
&\leq |(\hat{\Gamma}_n - \Gamma_n)\phi|_\infty + |\hat{\gamma}(n) - \gamma(n)|_\infty \\
&\leq |\hat{\Gamma}_n - \Gamma_n|_\infty |\phi|_1 + \epsilon \\
&\leq (|\phi|_1 + 1)\epsilon.
\end{aligned}$$

Choose $\lambda \geq (|\phi|_1+1)\epsilon$. Clearly ϕ is feasible for the SFSO for such λ and therefore $|\hat{\phi}|_1 \leq |\phi|_1$. Then

$$\begin{aligned}
|\hat{\phi}-\phi|_\infty &= |\Gamma_n^{-1}(\Gamma_n\hat{\phi}-\gamma(n))|_\infty \\
&\leq |\Gamma_n^{-1}|_{L^\infty}\left(|(\Gamma_n-\hat{\Gamma}_n)\hat{\phi}|_\infty + |\hat{\Gamma}_n\hat{\phi}-\hat{\gamma}(n)|_\infty + |\hat{\gamma}(n)-\gamma(n)|_\infty\right) \\
&\leq |\Gamma_n^{-1}|_{L^\infty}\left(|\Gamma_n-\hat{\Gamma}_n|_\infty|\hat{\phi}|_1 + |\hat{\Gamma}_n\hat{\phi}-\hat{\gamma}(n)|_\infty + |\hat{\gamma}(n)-\gamma(n)|_\infty\right) \\
&\leq |\Gamma_n^{-1}|_{L^\infty}\left(\epsilon|\phi|_1+\lambda+\epsilon\right) \\
&\leq 2\lambda|\Gamma_n^{-1}|_{L^\infty}.
\end{aligned}$$

Recall that $\hat{\gamma}(n) = (\hat{\gamma}_1,\ldots,\hat{\gamma}_n)^\top$, where $\hat{\gamma}_s = \kappa(|s|/l)\check{\gamma}_s$ and $\check{\gamma}_s = n^{-1}\sum_{t=1}^{n-|s|} X_tX_{t+|s|}$. Note that

$$\begin{aligned}
|\hat{\gamma}(n)-\hat{\gamma}(n)|_\infty &= \max_{1\leq s\leq n}|\kappa(s/l)\check{\gamma}_s-\gamma_s| \\
&\leq \max_{1\leq s\leq l}|\check{\gamma}_s-\gamma_s| + \max_{l<s\leq\lfloor c_\kappa l\rfloor}|\kappa(s/l)\check{\gamma}_s-\gamma_s| + \max_{\lfloor c_\kappa l\rfloor+1\leq s\leq n}|\gamma_s| \\
&:= T_1+T_2+T_3.
\end{aligned}$$

First, we deal with T_1. Observe that

$$\begin{aligned}
T_1 &\leq \max_{1\leq s\leq l}|\check{\gamma}_s-\mathbb{E}\check{\gamma}_s| + \max_{1\leq s\leq l}|\mathbb{E}\check{\gamma}_s-\gamma_s| \\
&= \frac{1}{n}\max_{1\leq s\leq l}|\sum_{t=1}^{n-s}(X_tX_{t+s}-\mathbb{E}X_tX_{t+s})| + \frac{1}{n}\max_{1\leq s\leq l} s|\gamma_s|.
\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
\|X_tX_{t+m}-X_t'X_{t+m}'\|_q &\leq \|X_t(X_{t+m}-X_{t+m}')\|_q + \|X_{t+m}'(X_t-X_t')\|_q \\
&\leq \|X_t\|_{2q}\|X_{t+m}-X_{t+m}'\|_{2q} + \|X_{t+m}'\|_{2q}\|X_t-X_t'\|_{2q}.
\end{aligned}$$

Since $\mathbb{E}|X_t|^{2q}\leq\nu$ and $\delta_{2q,t}\leq C_qt^{-d-1}$, it follows that the functional dependence measure of the process (X_tX_{t+m}) is bounded by $2C_q\nu^{1/2q}t^{-d-1}$ for all $m\geq 0$. For $d>1/2-1/q$, by the Nagaev inequality [LXW13], we have for all $\epsilon>0$

$$\mathbb{P}\left(\max_{1\leq s\leq l}|\sum_{t=1}^{n-s}(X_tX_{t+s}-\mathbb{E}X_tX_{t+s})|\geq\epsilon\right)\leq C_1l\left[\frac{n-s}{\epsilon^q}+\exp\left(-\frac{C_2\epsilon^2}{(n-s)^2}\right)\right].$$

Therefore, it follows that

$$T_1 = O_P(\epsilon^* + n^{-1}\max_{1\leq s\leq l}s|\gamma_s|),$$

where

$$\epsilon^* = \max\left\{\frac{l^{1/q}}{n^{1-1/q}}, \sqrt{\frac{\log l}{n}}\right\}.$$

For T_2, we note that

$$\begin{aligned} T_2 &\le \max_{l<s\le\lfloor c_\kappa l\rfloor} \kappa(s/l)|\check{\gamma}_s - \gamma_s| + \max_{l<s\le\lfloor c_\kappa l\rfloor} |\kappa(s/l)-1||\gamma_s| \\ &\le \max_{l<s\le\lfloor c_\kappa l\rfloor} |\check{\gamma}_s - \gamma_s| + \max_{l<s\le\lfloor c_\kappa l\rfloor} |\gamma_s|. \end{aligned}$$

Since $c_\kappa \ge 1$ is a constant, by the Nagaev inequality [LXW13], we have

$$T_2 = O_P\left(\epsilon^* + n^{-1}\max_{l<s\le\lfloor c_\kappa l\rfloor} s|\gamma_s| + \max_{l<s\le\lfloor c_\kappa l\rfloor}|\gamma_s|\right).$$

Combining the three terms together, we therefore have

$$|\hat{\gamma}(n) - \hat{\gamma}(n)|_\infty = O_P\left(\epsilon^{**}\right), \quad \text{where } \epsilon^{**} = \epsilon^* + n^{-1}\max_{1\le s\le\lfloor c_\kappa l\rfloor} s|\gamma_s| + \max_{l<s\le n}|\gamma_s|.$$

Since $\hat{\Gamma}_n$ and Γ_n are both Toeplitz matrices, the same bound applies for $|\hat{\Gamma}_n - \Gamma_n|_\infty$. Therefore for $\lambda \ge (|\phi|_1+1)\epsilon^{**}$ we get

$$|\hat{\phi} - \phi|_\infty = O_P(|\Gamma_n^{-1}|_{L^\infty}\lambda).$$

Let $u \ge 0$ and $D(u) = \sum_{i=1}^n \min(|\phi_i|, u)$ be the smallness measure of ϕ defined in [CXW13]. By the argument in [CXW15], $|\hat{\phi} - \phi|_1 \le 6D(3|\hat{\phi} - \phi|_\infty)$ and therefore by interpolation we have that

$$|\hat{\phi} - \phi|_2^2 = O_P\left(|\Gamma_n^{-1}|_{L^\infty}\lambda D(3|\Gamma_n^{-1}|_{L^\infty}\lambda\right).$$

Since $\phi \in \mathcal{G}_r(C_0, M)$, it is easy to see that $D(u) \le 2Mu^{1-r}$. Hence, we obtain

$$|\hat{\phi} - \phi|_2 = O_P(M^{1/2}|\Gamma_n^{-1}|_{L^\infty}^{1-r/2}\lambda^{1-r/2}). \qquad \square$$

References

[CLL11] Cai, T., Liu, W., and Luo, X., A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011. MR2847973

[CT07] Candès, E. and Tao, T., The dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007. MR2382644

[CXW13] Chen, X., Xu, M., and Wu, W. B., Covariance and precision matrix estimation for high-dimensional time series. *Annals of Statistics*, 41(6):2994–3021, 2013. MR3161455

[CXW15] Chen, X., Xu, M., and Wu, W. B., Estimation of covariance matrices and their functionals for high-dimensional linear processes. *Preprint*, 2015+.

[LXW13] Liu, W., Xiao, H., and Wu, W. B., Probability and moment inequalities under dependence. *Statistica Sinica*, 23:1257–1272, 2013. MR3114713

[MP10] McMurry, T. L. and Politis, D. N., Banded and tapered estimates for autocovairance matrices and the linear bootstrap. *Journal of Time Series Analysis*, 31:471–482, 2010. MR2732601

[MP15] McMurry, T. L. and Politis, D. N., High-dimensional autocovariance matrices and optimal linear prediction. *Electronic Journal of Statistics*, 9:753–788, 2015.