INFERENCE OF HIGH-DIMENSIONAL LINEAR MODELS WITH TIME-VARYING COEFFICIENTS

University of Illinois at Urbana-Champaign

Supplementary Material

The supplementary material contains additional technical lemmas and discusses some implementation issues.

Additional technical lemmas

Lemma S0.1. *Let X be an $n \times p$ matrix and $D = diag(d_1, \cdots, d_n)$ with $|d_i| \le b$ and $b \ge 0$. Then*

$$\rho_{\max}(X^\top D X, s) \le 2b\rho_{\max}(X^\top X, s).$$

If $d_i \in [0, b]$, then $\rho_{\max}(X^\top D X, s) \le b\rho_{\max}(X^\top X, s)$.

Proof. Let $\mathcal{A}_s = \{\mathbf{a} \in \mathbb{R}^p : |\mathbf{a}|_2 \le 1, |\mathbf{a}|_0 \le s\}$. Write $d_i = d_i^+ - d_i^-$, where $d_i^+ = \max(d_i, 0)$ and $d_i^- = \max(-d_i, 0)$ are the positive and negative parts, respectively. By definition

$$\begin{aligned}\rho_{\max}(X^\top DX,s) &= \max_{\mathbf{a}\in\mathcal{A}_s}|\mathbf{a}^\top X^\top DX\mathbf{a}| = \max_{\mathbf{a}\in\mathcal{A}_s}|\mathrm{tr}(D(X\mathbf{a}\mathbf{a}^\top X^\top))| \\ &= \max_{\mathbf{a}\in\mathcal{A}_s}\left|\sum_{i=1}^n (d_i^+ - d_i^-)(X\mathbf{a}\mathbf{a}^\top X^\top)_{ii}\right| \le 2b\max_{\mathbf{a}\in\mathcal{A}_s}\sum_{i=1}^n (X\mathbf{a}\mathbf{a}^\top X^\top)_{ii} \\ &= 2b\max_{\mathbf{a}\in\mathcal{A}_s}\mathrm{tr}(X\mathbf{a}\mathbf{a}^\top X^\top) = 2b\max_{\mathbf{a}\in\mathcal{A}_s}\mathbf{a}^\top X^\top X\mathbf{a} = 2b\rho_{\max}(X^\top X,s),\end{aligned}$$

because $X^\top X$ is nonnegative definite. The second claim follows from the same lines with $d_i^- = 0$. □

Lemma S0.2. *Let $t\in\varpi$ and $\hat{\Sigma}_t$ be the kernel smoothed sample covariance at time t and $\hat{\Sigma}_t^\diamond = \mathcal{X}_t^{\diamond\top}\mathcal{X}_t^\diamond$. Suppose that $\mathcal{X}_t^\diamond$ has full row rank. Assume further (15), (13) and assumption 6 hold, then we have*

$$\rho_{\min\neq 0}(\hat{\Sigma}_t) \ \ge\ |N_t|\underline{w}_t\varepsilon_0^2, \tag{S0.1}$$

$$\rho_{\max}(\hat{\Sigma}_t,s) \ \le\ |N_t|\overline{w}_t\varepsilon_0^{-2}. \tag{S0.2}$$

Proof. Since $\mathcal{X}_t^\diamond$ is of full row rank, $r=|N_t|$. Note that $\mathcal{X}_t = (|N_t|W_t)^{1/2}\mathcal{X}_t^\diamond$, $\rho_i(\hat{\Sigma}_t) = \sigma_i^2(\mathcal{X}_t)$ and $\rho_i(\hat{\Sigma}_t^\diamond) = \sigma_i^2(\mathcal{X}_t^\diamond)$. By the generalized Marshall-Olkin inequality, see e.g. (Wang and Zhang, 1992, Theorem 4), assumption 6 and (15), we have

$$\begin{aligned}\rho_{\min\neq 0}(\hat{\Sigma}_t) &= \rho_{\min}(\mathcal{X}_t\mathcal{X}_t^\top) = |N_t|\rho_{\min}(W_t^{1/2}\mathcal{X}_t^\diamond\mathcal{X}_t^{\diamond\top}W_t^{1/2}) \\ &= |N_t|\rho_{\min}(\mathcal{X}_t^\diamond\mathcal{X}_t^{\diamond\top}W_t) \ge |N_t|\rho_{\min}(W_t)\rho_{\min}(\mathcal{X}_t^\diamond\mathcal{X}_t^{\diamond\top}) \ge |N_t|\underline{w}_t\varepsilon_0^2.\end{aligned}$$

The second inequality (S0.2) follows from assumption 3(b) and Lemma S0.1 applying to $\hat{\Sigma}_t = |N_t| \mathcal{X}_t^{\diamond\top} W_t \mathcal{X}_t^{\diamond}$ and $W_t \geq 0$. □

Lemma S0.3. *Suppose assumption 1, 2, 3 and 5(a) hold. Let* $t \in \varpi$ *be fixed and* λ_0 *be defined in (17). Then, for* $\lambda_1 \geq 2(\lambda_0 + 2C_0 L_{t,1} s^{1/2} \varepsilon_0^{-1} b_n |N_t| \overline{w}_t)$ *where* λ_0 *is defined in (17), we have, with probability* $1 - 2p^{-1}$,

$$|\mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]|_2^2 + \lambda_1 |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 \leq 4\lambda_1^2 \frac{s}{\phi_0^2}. \tag{S0.3}$$

Proof. By definition (8),

$$|\mathcal{Y}_t - \mathcal{X}_t \tilde{\boldsymbol{\beta}}(t)|_2^2 + \lambda_1 |\tilde{\boldsymbol{\beta}}(t)|_1 \leq |\mathcal{Y}_t - \mathcal{X}_t \boldsymbol{\beta}(t)|_2^2 + \lambda_1 |\boldsymbol{\beta}(t)|_1,$$

which implies that

$$|\mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]|_2^2 + \lambda_1 |\tilde{\boldsymbol{\beta}}(t)|_1 \leq \lambda_1 |\boldsymbol{\beta}(t)|_1 + 2\left\langle \mathcal{Y}_t - \mathcal{X}_t \boldsymbol{\beta}(t), \mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)] \right\rangle.$$

By assumption 2 and Taylor's expansion in the b_n-neighborhood of t, we see that

$$\mathcal{Y}_t - \mathcal{X}_t \boldsymbol{\beta}(t) = \mathcal{E}_t + M_t \mathcal{X}_t \boldsymbol{\beta}'(t) + \mathcal{X}_t \boldsymbol{\xi}, \tag{S0.4}$$

where $M_t = \operatorname{diag}((t_i - t)_{i \in N_t})$ and $\boldsymbol{\xi}$ is a vector such that $|\boldsymbol{\xi}|_\infty \leq C_0 b_n^2/2$ and $|\boldsymbol{\xi}|_0 \leq s$. Let $\mathcal{J} = \{2|\mathcal{E}_t^\top \mathcal{X}_t|_\infty \leq \lambda_0\}$. Observe that $|\mathcal{E}_t^\top \mathcal{X}_t|_\infty = \max_{j \leq p} |\sum_{i \in N_t} w(i,t) X_{ij} e_i|$ and, by assumption 1,

$$\sum_{i \in N_t} w(i,t) X_{ij} e_i \sim N\left(0, \sigma^2 \sum_{i \in N_t} w(i,t)^2 X_{ij}^2\right). \tag{S0.5}$$

Then, by the standard Gaussian tail bound and the union bound, we obtain that

$$\mathbb{P}\left(\max_{j\le p}\left|\frac{\sum_{i\in N_t} w(i,t)X_{ij}e_i}{\sigma L_{t,2}}\right| \ge \sqrt{\varepsilon^2+2\log p}\right) \le \mathbb{P}(\max_{j\le p}|Z_j| \ge \sqrt{\varepsilon^2+2\log p}) \le 2\exp\left(-\frac{\varepsilon^2}{2}\right)$$

for all $\varepsilon > 0$, where $Z_j \sim N(0,1)$. Now, choose $\varepsilon = (2\log p)^{1/2}$ and $\lambda_0 = 4\sigma L_{t,2}(\log p)^{1/2}$, we have $\mathbb{P}(\mathcal{J}) \ge 1 - 2p^{-1}$. Further, we have

$$\begin{aligned}
& |\boldsymbol{\beta}'(t)^\top \mathcal{X}_t^\top M_t \mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]| \le |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 |\mathcal{X}_t^\top M_t \mathcal{X}_t \boldsymbol{\beta}'(t)|_\infty \\
\le\ & |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 \max_{j\le p}\left(\sum_{i\in N_t} w(i,t)X_{ij}^2\right)^{1/2}\left[\boldsymbol{\beta}'(t)^\top \mathcal{X}_t^\top M_t^2 \mathcal{X}_t \boldsymbol{\beta}'(t)\right]^{1/2} \quad \text{(Cauchy-Schwarz)} \\
\le\ & |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 L_{t,1}\sqrt{\rho_{\max}(\mathcal{X}_t^\top M_t^2 \mathcal{X}_t, s)}|\boldsymbol{\beta}'(t)|_2 \quad \text{(assumption 2)} \\
\le\ & |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 L_{t,1} C_0 s^{1/2} b_n \sqrt{\rho_{\max}(\mathcal{X}_t^\top \mathcal{X}_t, s)} \quad \text{(Lemma S0.1, assumption 2 and 3)} \\
\le\ & |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 L_{t,1} C_0 (|N_t|\overline{w}_t s)^{1/2} b_n \varepsilon_0^{-1} \quad \text{(Lemma S0.2, equation (S0.2))}.
\end{aligned}$$

Similarly, we can show that $|\boldsymbol{\xi}^\top \mathcal{X}_t^\top \mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]| = O(L_{t,1}(|N_t|\overline{w}_t s)^{1/2} b_n^2 \varepsilon_0^{-1}|\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1)$. Therefore, it follows that, with probability at least $(1 - 2p^{-1})$,

$$\left|\left\langle \mathcal{Y}_t - \mathcal{X}_t\boldsymbol{\beta}(t), \mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]\right\rangle\right| \le \left[\lambda_0 + 2L_{t,1}C_0(|N_t|\overline{w}_t s)^{1/2} b_n \varepsilon_0^{-1}(1 + o(1))\right] |\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1.$$

Now, choose $\lambda_1 \ge 2(\lambda_0 + 2L_{t,1}C_0(|N_t|\overline{w}_t s)^{1/2} b_n \varepsilon_0^{-1})$, we get

$$2|\mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]|_2^2 + 2\lambda_1|\tilde{\boldsymbol{\beta}}(t)|_1 \le \lambda_1|\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)|_1 + 2\lambda_1|\boldsymbol{\beta}(t)|_1.$$

Denote $S_0 := S_0(t) = \mathrm{supp}(\boldsymbol{\beta}(t))$. By the same argument as (Bühlmann and van de Geer, 2011, Lemma 6.3), it is easy to see that, on $\mathcal{J}$,

$$2|\mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]|_2^2 + \lambda_1|\tilde{\boldsymbol{\beta}}_{S_0^c}(t)|_1 \le 3\lambda_1|\tilde{\boldsymbol{\beta}}_{S_0}(t) - \boldsymbol{\beta}_{S_0}(t)|_1.$$

But then, (S0.3) follows from the restricted eigenvalue condition (assumption 5) with the elementary inequality $4ab \le a^2 + 4b^2$ that

$$2|\mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}(t)]|_2^2+\lambda_1|\tilde{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}(t)|_1 \le 4\lambda_1|\tilde{\boldsymbol{\beta}}_{S_0}(t)-\boldsymbol{\beta}_{S_0}(t)|_1 \le |\mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}(t)]|_2^2+4\lambda_1^2 s/\phi_0^2.$$

□

Definition S0.1. A mean zero random variable is said to be *sub-Gaussian* with variance factor σ^2 if

$$\log \mathbb{E}(e^{\lambda X}) \le \lambda^2\sigma^2/2 \qquad \text{for all } \lambda \in \mathbb{R}.$$

Lemma S0.4. *Let ξ_i be iid sub-Gaussian random variables with mean zero and variance factor σ^2, and $e_i = \sum_{m=0}^{\infty} a_m \xi_{i-m}$ be a linear process. Let $\mathbf{w} = (w_1, \cdots, w_n)$ be a real vector and $S_n = \sum_{i=1}^{n} w_i e_i$ be the weighted partial sum of e_i.*

1. *(Short-range dependence). If $|\mathbf{a}|_1 = \sum_{i=0}^{\infty} |a_i| < \infty$, then for all $x > 0$ we have*
$$\mathbb{P}(|S_n| \ge x) \le 2\exp\left(-\frac{x^2}{2|\mathbf{w}|_2^2|\mathbf{a}|_1^2\sigma^2}\right). \tag{S0.6}$$

2. *(Long-range dependence). Suppose $K = \sup_{m\ge 0} |a_m|(m+1)^{\varrho} < \infty$, where $1/2 < \varrho < 1$. Then, there exists a constant C_ϱ that only depends on ϱ such that*
$$\mathbb{P}(|S_n| \ge x) \le 2\exp\left(-\frac{C_\varrho x^2}{|\mathbf{w}|_2^2 n^{2(1-\varrho)}\sigma^2 K^2}\right). \tag{S0.7}$$

Proof. Put $a_m = 0$ if $m < 0$ and we may write $S_n = \sum_{m\in\mathbb{Z}} b_m \xi_m$, where $b_m = \sum_{i=1}^n w_i a_{i-m}$. By the Cauchy-Schwarz inequality,

$$\sum_{m\in\mathbb{Z}} b_m^2 \le \sum_{m\in\mathbb{Z}} \left(\sum_{i=1}^n w_i^2 |a_{i-m}|\right)\left(\sum_{i=1}^n |a_{i-m}|\right) \le |\mathbf{w}|_2^2 |\mathbf{a}|_1^2.$$

Then, (S0.6) follows from the Cramér-Chernoff bound Boucheron et al. (2013). Let $\bar{a}_m = \max_{l\ge m} |a_l|$ and $A_m = \sum_{l=0}^m |a_l|$. Note that $A_n \le K\sum_{l=0}^n (l+1)^{-\varrho} \le C_\varrho K(n+1)^{1-\varrho}$, where $C_\varrho = (1-\varrho)^{-1}$. Then, we have

$$\sum_{m=1-n}^n b_m^2 \le \sum_{m=1-n}^n \left(\sum_{i=1}^n w_i^2 |a_{i-m}|\right)\left(\sum_{i=1}^n |a_{i-m}|\right) \le |\mathbf{w}|_2^2 A_{2n}^2.$$

If $m \le -n$, then $|b_m| \le |\mathbf{w}|_1 \bar{a}_{1-m}$ and therefore

$$\sum_{m\le -n} b_m^2 \le |\mathbf{w}|_1^2 \sum_{m\le -n} \bar{a}_{1-m}^2 \le C_\varrho n |\mathbf{w}|_2^2 K^2 n^{1-2\varrho},$$

where the last inequality follows from Karamata's theorem; see e.g. Resnick (1987). Hence, the proof is complete by invoking the Cramér-Chernoff bound for sub-Gaussian random variables. □

Some implementation issues

We assumed that the noise variance-covariance matrix Σ_e is known. In the iid error case $\Sigma_e = \sigma^2 I_n$, we have seen that the distribution $F(\cdot)$ is independent of σ^2 and therefore its value does not affect the inference procedure.

The noise variance only impacts the tuning parameter of the initial Lasso estimator. In practice, we can use the scaled Lasso to estimate σ^2 in our numeric studies. Given that $|\hat{\sigma}/\sigma - 1| = o_{\mathbb{P}}(1)$ Sun and Zhang (2012), the theoretical properties of our estimator (10) remains the same if we plug in the scaled Lasso variance output to our method. For temporally dependent stationary error process, estimation of Σ_e becomes more subtle since it involves n autocovariance parameters. We propose a heuristic strategy: first, run the tv-Lasso estimator and get the residuals; then calculate the sample autocovariance matrix and apply a banding or tapering operation $B_h(\Sigma) = \{\sigma_{jk}\mathbf{1}(|j-k| \le h)\}_{j,k=1}^p$ Bickel and Levina (2008); Cai et al. (2010); McMurry and Politis (2010).

We provide some justification on the heuristic strategy for SRD time series models. To simplify explanation, we consider the uniform kernel and the bandwidth $b_n = 1$. Suppose we have an oracle where $\boldsymbol{\beta}(t)$ is known and we have access to the error process $e(t)$. Let Σ_e^* be the oracle sample covariance matrix of e_i with the Toeplitz structure i.e. the h-th subdiagonal of Σ_e^* is $\sigma_{e,h}^* = n^{-1}\sum_{i=1}^{n-h} e_i e_{i+h}$. We first compare the oracle estimator and the true error covariance matrix Σ_e. Let $\alpha > 0$ and define

$$\mathcal{T}(\alpha, C_1, C_2) = \left\{ M \in ST^{p\times p} : \sum_{k=h+1}^{p} |m_k| \le C_1 h^{-\alpha}, \rho_j(M) \in [C_2, C_2^{-1}],\ \forall j = 1, \cdots, p \right\},$$

where $ST^{p\times p}$ is the set of all $p \times p$ symmetric Toeplitz matrices. If e_i has

SRD, then $\Sigma_e \in \mathcal{T}(\varrho - 1, C_1, C_2)$. By the argument in Bickel and Levina (2008) and Lemma S0.4, we can show that

$$\begin{aligned}\rho_{\max}(B_h(\Sigma_e^*) - \Sigma_e) &\leq \rho_{\max}(B_h(\Sigma_e^*) - B_h(\Sigma_e)) + \rho_{\max}(B_h(\Sigma_e) - \Sigma_e) \\ &\lesssim_{\mathbb{P}} h\sqrt{\frac{\log h}{n}} + h^{-(\varrho-1)}.\end{aligned}$$

Choosing $h^* \asymp (n/\log n)^{1/(2\varrho)}$, we get

$$\rho_{\max}(B_h(\Sigma_e^*) - \Sigma_e) = O_{\mathbb{P}}\left(\left(\frac{\log n}{n}\right)^{\frac{\varrho-1}{2\varrho}}\right).$$

This oracle rate is sharper than the one established in Bickel and Levina (2008) for regularizing more general bandable matrices if $n = o(p)$. Here, the improved rate is due to the Toeplitz structure in Σ_e. Since Σ_e has uniformly bounded eigenvalues from zero and infinity, the banded oracle estimator $B_h(\Sigma_e^*)$ can be used as a benchmark to assess the tv-Lasso residuals $\tilde{\mathcal{E}}_t = \mathcal{Y}_t - \mathcal{X}_t\tilde{\boldsymbol{\beta}}(t)$.

Proposition S0.5. *Suppose $\Sigma_e \in \mathcal{T}(\varrho-1, C_1, C_2)$ and conditions of Lemma S0.3 are satisfied except that (e_i) is an SRD stationary Gaussian process with $\varrho > 1$. Then*

$$\rho_{\max}(B_h(\hat{\Sigma}_e) - B_h(\Sigma_e^*)) = O_{\mathbb{P}}(h\lambda_1 s^{1/2}). \tag{S0.8}$$

With the choice $h^* \asymp (n'/\log n')^{1/2\varrho}$ where $n' = |N_t|$, we have

$$\rho_{\max}(B_h(\hat{\Sigma}_e) - \Sigma_e)) = O_{\mathbb{P}}\left(\left(\frac{\log n'}{n'}\right)^{\frac{\varrho-1}{2\varrho}} + \left(\frac{n'}{\log n'}\right)^{\frac{1}{2\varrho}}\left(\sqrt{\frac{s\log p}{n'}} + sb_n)\right)\right). \tag{S0.9}$$

It is interesting to note that the price we pay to choose h for not knowing the error process is the second term in (S0.9). Bandwidth selection for the smoothing parameter b_n is a theoretically challenging task in the high dimension. Asymptotic optimal order for the parameter is available up to some unknown constants depending on the data generation parameters. We shall use the cross-validation (CV) in our simulation studies and real data analysis.

Proof of Proposition S0.5. Since we consider the uniform kernel, we may assume $b_n = 1, |N_t| = n$ and then rescale. Observe that

$$\begin{aligned}
\max_{|k|\le h} |\hat{\sigma}^2_{e,k} - \sigma^{*2}_{e,k}| &= \max_{|k|\le h} \frac{1}{n} \left| \sum_{i=1}^{n-k} (\hat{e}_i \hat{e}_{i+k} - e_i e_{i+k}) \right| \\
&\le \max_{|k|\le h} \frac{1}{n} \left| \sum_{i=1}^{n-k} \hat{e}_i (\hat{e}_{i+k} - e_{i+k}) \right| + \left| \sum_{i=1}^{n-k} e_{i+k} (\hat{e}_i - e_i) \right| \\
&\le \max_{|k|\le h} \frac{1}{n} \left(\sum_{i=1}^{n-k} \hat{e}_i^2 \right)^{1/2} \left(\sum_{i=1}^{n-k} (\hat{e}_{i+k} - e_{i+k})^2 \right)^{1/2} \\
&\quad + \max_{|k|\le h} \frac{1}{n} \left(\sum_{i=1}^{n-k} e_{i+k}^2 \right)^{1/2} \left(\sum_{i=1}^{n-k} (\hat{e}_i - e_i)^2 \right)^{1/2} \\
&\le \left[\left(\frac{1}{n} \sum_{i=1}^{n} \hat{e}_i^2 \right)^{1/2} + \left(\frac{1}{n} \sum_{i=1}^{n} e_i^2 \right)^{1/2} \right] \left(\frac{1}{n} \sum_{i=1}^{n} (\hat{e}_i - e_i)^2 \right)^{1/2}.
\end{aligned}$$

By Lemma S0.3,

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{e}_i - e_i)^2 = |\tilde{\mathcal{E}}_t - \mathcal{E}_t|_2^2 = |\mathcal{X}_t[\tilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)]|_2^2 = O_{\mathbb{P}}(\lambda_1^2 s).$$

Then, it follows from the last expression and $n^{-1}\sum_{i=1}^{n} e_i^2 = O_{\mathbb{P}}(1)$ that

$$\max_{|k|\le h} |\hat{\sigma}_{e,k}^2 - \sigma_{e,k}^{*2}| = O_{\mathbb{P}}(\lambda_1 s^{1/2}).$$

Therefore

$$\rho_{\max}(B_h(\hat{\Sigma}_e) - B_h(\Sigma_e^*)) \lesssim h \max_{|k|\le h} |\hat{\sigma}_{e,k}^2 - \sigma_{e,k}^{*2}| = O_{\mathbb{P}}(h\lambda_1 s^{1/2}).$$

□

References

Bickel, P. J. and Levina, E. (2008). Regularized Estimation of Large Covariance Matrices, *The Annals of Statistics* **36**(1): 199–227.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Series in Statistics.

Cai, T., Zhang, C.-H. and Zhou, H. (2010). Optimal Rates of Convergence for Covariance Matrix Estimation, *The Annals of Statistics* **38**(4): 2118–2144.

McMurry, T. L. and Politis, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap., *J. Time Ser. Anal.* **31**: 471–482.

Resnick, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes*, Applied Probability, Springer-Verlag.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression, *Biometrika* **99**: 879–898.

Wang, B. and Zhang, F. (1992). Some inequalities for the eigenvalues of the product of positive semidefinite Hermitian matrices, *Linear Algebra and Its Applications* **160**: 113–118.