**Supplementary materials for the paper "Nonparametric Eigenvalue-Regularized Precision or Covariance Matrix Estimator"- Simulation results and the proof of Theorem 1, Theorem 3, Lemma 1, Theorem 5 and Theorem 6 in the paper.**

# 1 Simulation Experiments

We demonstrate and compare the performance of our estimator to other state-of-the-art estimators under various settings. Hereafter, we abbreviate our method as NERCOME for estimating $\mathbf{\Sigma}_p$ or $\mathbf{\Omega}_p$, which comes from the name Nonparametric Eigenvalue-Regularized COvariance Matrix Estimator. The method proposed in Abadir et al. (2014) is abbreviated as CRC (Condition number Regularized Covariance estimator), while the nonlinear shrinkage method in Ledoit and Wolf (2012) is abbreviated as NONLIN. We call the grand average estimator (15) in Abadir et al. (2014) the CRC grand average. Finally, the method in Fan et al. (2013) is abbreviated as POET. We also include in some cases the graphical LASSO, abbreviated as GLASSO, as in Friedman et al. (2008), and the adaptive SCAD thresholding, abbreviated as SCAD, which is a special case of POET without any factors. We create five different profiles:

(I) Independent data $\mathbf{y}_t \sim N(\mathbf{0}, \mathbf{\Sigma}_p)$, where $\mathbf{\Sigma}_p = \mathbf{QDQ}^{\mathrm{T}}$. The orthogonal matrix $\mathbf{Q}$ is randomly generated each time, and $\mathbf{D}$ is diagonal with 40% of values being 3, and 60% being 7.

(II) Same as (I), except that $\mathbf{\Sigma}_p = \mathbf{I}_p$.

(III) The data is from a factor model $\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \epsilon_t$, where $\mathbf{A}$ has size $p \times 3$, with elements in $\mathbf{A}$ generated independently from $N(0, 2^2)$. The $\mathbf{x}_t$'s are independently generated from $N(\mathbf{0}, 2\mathbf{I}_r)$, while the $\boldsymbol{\epsilon}_t$'s are independently generated from the standardized $t_3$ distribution.

(IV) (Sparse covariance matrix). The covariance matrix $\mathbf{\Sigma}_p$ is sparse and is randomly generated each time, with 20% non-zeros, and independent data $\mathbf{y}_t \sim N(\mathbf{0}, \mathbf{\Sigma}_p)$.

(V) Same as (IV), except that $\mathbf{y}_t = \mathbf{\Sigma}_p \mathbf{z}_t$ where the $\mathbf{z}_t$'s are independent containing independent standardized $t_5$ random variables.

Profiles (I), (II), (IV) and (V) are for non-factor model, with profile (II) being excluded in Theorem 5, since the loss of an ideal estimator is exactly 0 when $\mathbf{\Sigma}_p = \sigma^2 \mathbf{I}_p$. Profile (III) is for factor model. Profiles (III) and (V) test the robustness of our method to fat-tailed distributions, since assumption (F1) rules out an error distribution of $t_3$, while assumption (A1) rules out $t_5$ being the distribution of the independent random variables.

Recall that NONLIN is proved to be asymptotically optimal with respect to the Frobenius and the inverse Stein's loss functions when the data is not from a factor model (see the results in Ledoit and Wolf (2013a) for more details).

Hence we expect that NERCOME should be close to NONLIN in performance for the non-factor model profiles, but better than NONLIN for profile (III), since NERCOME is proved to be asymptotically optimal in Theorem 3 for data from a factor model.

We simulate 500 times from the seven profiles under all different combinations of $n = 200, 400, 800$ and $p = 50, 100, 200, 500$, and calculate the mean efficiency loss defined in (4.2) with respect to the Frobenius loss and the inverse Stein's loss given in (4.3) and (2.16) respectively. For profiles (IV) and (V), we include 5 more loss functions for comparisons. See Table 1 and the paragraphs therein for details. For both NERCOME and CRC, we use $M = 50$ permutations for each split of the data, which gives a good trade-off between accuracy and computational complexity. We use the 7 different split locations as in (4.8) for NERCOME and CRC. We also demonstrate the performance of NERCOME using the automatic choice of $m$ by minimizing $g(m)$ defined in (4.7). The CRC grand average estimator defined in (18) of Abadir et al. (2014) will also be included for comparison, which takes average over the 4 CRC estimators calculated on the split locations $0.2n, 0.4n, 0.6n$ and $0.8n$ respectively.

In the titles of the subsequent graphs, the inverse Stein's loss is abbreviated as the Stein's loss to save space.

Figure 1 shows the mean efficiency loss over 500 simulations for various methods under profile (I). Clearly, the graphical LASSO and the SCAD adaptive thresholding are not performing as good as other methods in terms of the Frobenius loss function, since the covariance and the precision matrices under profile (I) are only approximately sparse. The graphical LASSO has the worst performance, which is to be expected since it concerns with the sparse precision matrix rather than the covariance matrix estimation. NONLIN is the best in all scenarios, followed by NERCOME with $m$ chosen automatically by minimizing (4.7), and CRC grand average. It is fair to say that NONLIN, NERCOME and CRC grand average all perform well in absolute terms. Comparing the performance at different split locations for NERCOME and CRC, we can see that CRC attains a minimum with a smaller split $m$, and performs better when $m$ is not large. The opposite is true for NERCOME, which usually is at the smallest efficiency loss when $m$ is large. This is consistent with the condition $m/n \to 1$ for asymptotic efficiency in Theorem 5. Clearly, the choice of $m$ that minimizes $g(m)$ in (4.7) is good, since the resulting mean efficiency loss is close to the minimum for NERCOME. We can also see that for CRC, averaging over estimators with split locations from $0.2n$ to $0.8n$, which is the CRC grand average estimator suggested in Abadir et al. (2014), may result in suboptimal performance, since the CRC is usually suboptimal when $m$ is small.

We have omitted the mean efficiency loss with respect to the inverse Stein's loss for profile (I), since the graphs are very similar to the Frobenius loss's counterparts in Figure 1.

Figure 2 shows the actual Frobenius loss and inverse Stein's loss for various methods for profile (II). The efficiency loss in (4.2) is always 1 for any imperfect estimators, since it is easy to see that $\widehat{\mathbf{\Sigma}}_{\text{Ideal}} = \mathbf{\Sigma}_p = \sigma^2 \mathbf{I}_p$. Even for $\mathbf{\Sigma}_p = \sigma^2 \mathbf{I}_p$,

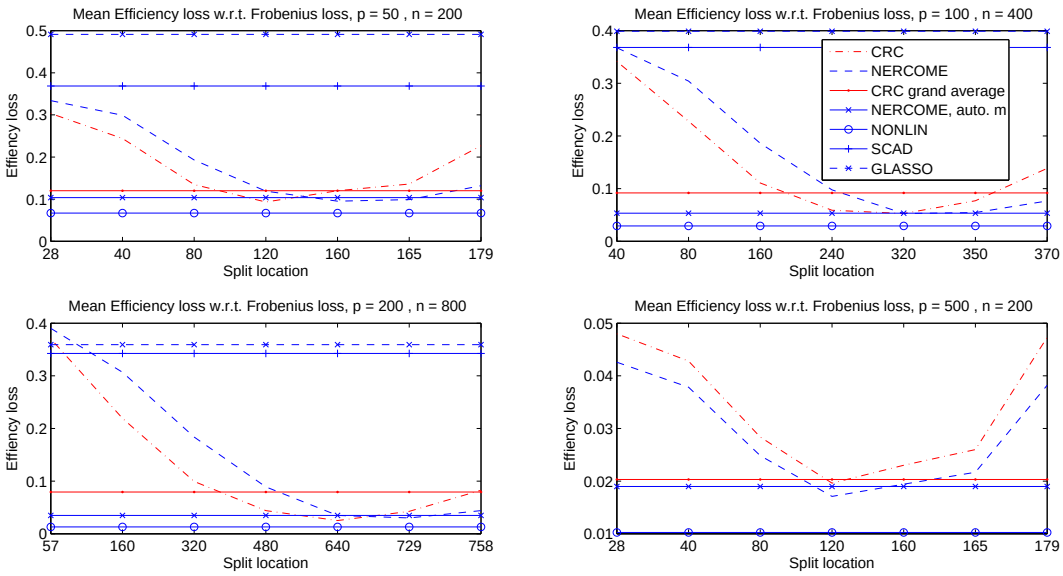


Figure 1: Mean efficiency loss with respect to the Frobenius loss for profile (I). SCAD and GLASSO are included only when $p < 500$ to save computational time. NERCOME with $m$ automatically chosen is to minimize $g(m)$ in (4.7) over the split locations in (4.8). CRC shares the same split locations. Only NERCOME and CRC are varying over different split locations.

SCAD thresholding and graphical LASSO are not as good as other methods in terms of both loss functions. This can be explained by the fact that even if all the off-diagonal elements of the covariance and the precision matrix estimated from the data are correctly killed off, the diagonal elements are left untouched. Hence, the errors in the diagonal elements will still stack up in the Frobenius or the inverse Stein's loss. On the other hand, CRC, NERCOME and NONLIN are in principal setting the eigenvalues to be as close to $\mathbf{p}_i^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_i$ as possible. Under $\mathbf{\Sigma}_p = \sigma^2\mathbf{I}_p$, $\mathbf{p}_i^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_i = \sigma^2$, which is the correct magnitude. The off-diagonal entries will then be close to 0 since $\mathbf{P}\mathrm{diag}(\sigma^2,\ldots,\sigma^2)\mathbf{P}^{\mathrm{T}} = \mathbf{P}(\sigma^2\mathbf{I}_p)\mathbf{P}^{\mathrm{T}} = \sigma^2\mathbf{I}_p$, no matter what orthogonal $\mathbf{P}$ we use. Hence, we can see that the advantage of CRC, NERCOME and NONLIN comes from the fact that the class of estimators $\mathbf{\Sigma}(\mathbf{D}) = \mathbf{PDP}^{\mathrm{T}}$ is particularly good when the true covariance matrix is diagonal, with the same entries on the main diagonal throughout. The actual performance of CRC, NERCOME and NONLIN are very close in this case.

It is clear that for both NERCOME and CRC, the best split location is when $m$ is the smallest. This can be seen for NERCOME by noting that, when $\mathbf{\Sigma}_p = \sigma^2\mathbf{I}_p$, $\left\|\widehat{\mathbf{\Sigma}}_m - \mathbf{\Sigma}_p\right\|_F^2 = \sum_{i=1}^p(\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i} - \sigma^2)^2$. Hence, in order to minimize

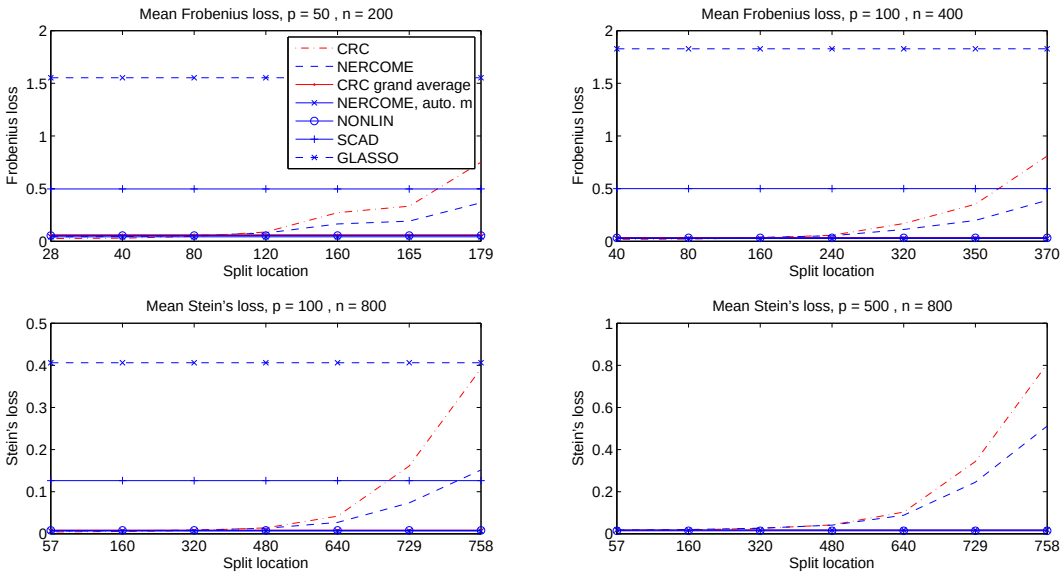


Figure 2: Mean Frobenius loss (upper row) and inverse Stein's loss (lower row) for profile (II). SCAD and GLASSO are absent for $p < 500$. Refer to the descriptions of Figure 1 for further details.

the loss, we want $\widetilde{\mathbf{\Sigma}}_2$ to be as close to $\sigma^2\mathbf{I}_p$ as possible. It means that $n_2 = n-m$ should be made as large as possible, which implies that $m$ should be as small as possible. The automatic selection for NERCOME clearly performs well even in this case.

For profile (III), all the methods perform badly in terms of Frobenius loss with high efficiency losses, hence the corresponding figure is not shown. This can be explained by the fact that since the factor model has 3 spiked eigenvalues of order $p$, any error in estimating those eigenvalues will be hugely reflected in the efficiency loss.

On the other hand, Figure 3 shows a huge improvement for CRC and NERCOME when the inverse Stein's loss is concerned, while POET performs the worst and NONLIN stays at around 50% efficiency loss. In one simulation run for $n = p = 200$, NONLIN returns a singular covariance matrix, and hence the inverse Stein's loss is undefined. This case has to be removed for calculating the mean of efficiency loss. Note that NONLIN always results in a positive semi-definite covariance matrix. However, when $n = p = 200$, it corresponds to $c = 1$ with exactly 1 zero eigenvalue for the sample covariance matrix. This case is not covered in Theorem 4 of Ledoit and Péché (2011), and the fact that the data is from a factor model violates the important assumption in Ledoit and Péché

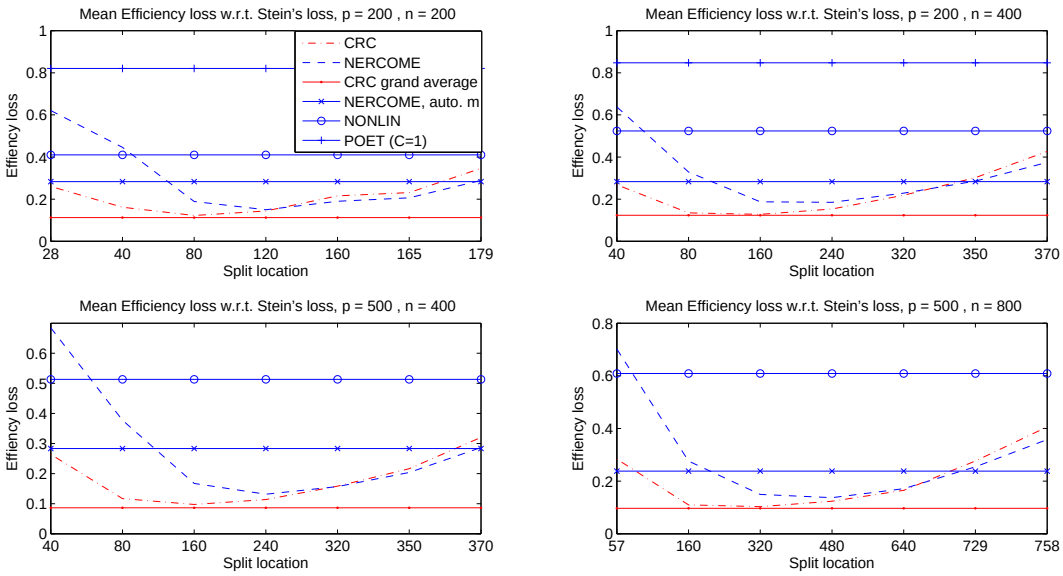


Figure 3: Mean efficiency loss with respect to the inverse Stein's loss for profile (III). The mean efficiency for NONLIN when $n = p = 200$ (upper left panel) is calculated after removing one simulation result with singular covariance matrix.

(2011) and Ledoit and Wolf (2013a,b), that we are able to write $\mathbf{y}_t = \mathbf{\Sigma}_p^{1/2}\mathbf{z}_t$, with $\mathbf{z}_t$ having independent and identically distributed entries. Hence, although rare in practice, it is not surprising to see a singular covariance matrix estimator for NONLIN. On the other hand, we have proved in Corollary 4 that our estimator is almost surely positive definite even for data from a factor model, although assumption (F1) is not satisfied here.

CRC grand average performs the best now, since it appears that all the split locations between $0.2n$ and $0.8n$ are indeed giving good performance to CRC and NERCOME. Part of the reason that POET is not as good as CRC, NERCOME and NONLIN is related to the fact that $\mathbf{\Sigma}_\epsilon = \mathbf{I}_p$, which gives advantages to the class of estimators $\mathbf{\Sigma}(\mathbf{D}) = \mathbf{PDP}^{\mathrm{T}}$, with arguments similar to the explanations of why SCAD thresholding and graphical LASSO perform worse under profile (II).

For profiles (IV) and (V), we investigate the performance of various estimators in estimating sparse covariance matrix. NONLIN, NERCOME and CRC are all rotation-equivariant estimators, and regularize through stabilizing the eigenvalues. Hence they are not recovering the sparse structure, unlike the thresholding estimator SCAD. On top of the two original loss functions, we

introduce 5 more to gauge the performance:

$$L_3(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) = \left\| \mathbf{\Sigma}^{-1} - \widehat{\mathbf{\Sigma}}^{-1} \right\|_F, \quad L_4(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) = \mathrm{tr}(\mathbf{\Sigma}^{-1}\widehat{\mathbf{\Sigma}}) - \log\det(\mathbf{\Sigma}^{-1}\widehat{\mathbf{\Sigma}}) - p,$$

$$L_5(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) = \left\| \mathbf{\Sigma} - \widehat{\mathbf{\Sigma}} \right\|, \qquad L_6(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) = \sum_{i=1}^{p} |\lambda_i(\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}})|,$$

$$L_7(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) = \mathrm{tr}(\mathbf{\Sigma} + \widehat{\mathbf{\Sigma}} - 2\mathbf{\Sigma}^{1/2}\widehat{\mathbf{\Sigma}}^{1/2}).$$

$L_3$ is the Frobenius norm on the difference of the inverse covariance matrices; $L_4$ is the Stein's loss; $L_5$ is the operator/spectral norm of the difference; $L_6$ is the nuclear norm of the difference. Finally, $L_7$ is called the Fréchet loss. We use $L_1$ and $L_2$ for the Frobenius and the inverse Stein's loss, respectively.

Table 1 shows the results for profile (IV) when $p = 200$. Other combinations of $(n, p)$ are omitted to save space. Sample covariance always leads to huge efficiency loss and hence is not shown. It is clear that when $\mathbf{\Sigma}_p$ is sparse, SCAD thresholding can take advantage and usually has negative efficiency loss when the dimension is not too close to the sample size. However, if the dimension is close to the sample size, SCAD has difficulty in finding a good thresholding parameter, resulting in large efficiency losses across different loss functions. NONLIN usually outperforms CRC grand average and NERCOME apart from SCAD, and efficiency loss is usually getting smaller for NONLIN when sample size gets larger while dimension stays fixed. NERCOME follows NONLIN closely in all $(p, n)$ combinations, and outperforms in several occasions, especially when dimension is close to the sample size. In fact NERCOME seems to do particularly well in the Frobenius loss for the inverse and the Stein's loss. CRC grand average, on the other hand, does not usually have decreasing efficiency loss as sample size increases. On the contrary, many actually increase as sample size increases while dimension stays fixed, showing that averaging on different splits is not always a good idea, especially when $n$ is large compared to $p$, which is a setting that Abadir et al. (2014) used. Figure 1 also shows this phenomenon. In all cases, we can see that among the three eigenvalues-stabilizing methods, NERCOME either outperforms, or is close to the best performer.

The same table shows the results for profile (V) when $p \geq 100$. With fat -tailed distribution, it is more difficult for SCAD to find a good thresholding parameter (except when $n$ is 8 or more times larger than $p$). CRC grand average still shows an increasing efficiency loss in general when $n$ is getting larger. NONLIN is now less dominant among the three eigenvalues-stabilizing methods, with NERCOME outperforming both of the other methods in more $(p, n)$ combinations. NONLIN has produced two singular covariance matrices in the $p = n = 200$ simulations, showing that when assumption (A1) is violated, NONLIN can in fact produce singular covariance matrix. NERCOME and CRC grand average still produce positive definite covariance matrices in those two cases. Again, among the three methods, NERCOME either outperforms or is close to the best performer in all $(p, n)$ settings.

Figure 4 shows the boxplots of the efficiency loss of the averaged ideal estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{Ideal},m,M} = M^{-1}\sum_{i=1}^{M} \widehat{\mathbf{\Sigma}}^{(i)}_{\mathrm{Ideal},m}$ for various profiles. We show this

| | | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | $L_7$ |
|---|---|---|---|---|---|---|---|---|
| Profile (IV) | | | | | | | | |
| $p = 200$, $n = 200$ | NERCOME | $3.8_{(1.0)}$ | $3.5_{(1.0)}$ | $\mathbf{0.7_{(1.3)}}$ | $\mathbf{1.0_{(1.6)}}$ | $5.2_{(1.8)}$ | $1.7_{(0.4)}$ | $2.5_{(0.6)}$ |
| | CRC | $3.0_{(0.5)}$ | $\mathbf{2.7_{(0.5)}}$ | $1.8_{(1.1)}$ | $1.7_{(1.4)}$ | $\mathbf{2.8_{(1.3)}}$ | $1.3_{(0.3)}$ | $2.5_{(0.5)}$ |
| | NONLIN | $\mathbf{2.6_{(1.8)}}$ | $3.5_{(6.7)}$ | $2.4_{(10.6)}$ | $1.2_{(3.2)}$ | $3.6_{(6.7)}$ | $\mathbf{1.2_{(0.7)}}$ | $\mathbf{2.3_{(2.6)}}$ |
| | SCAD | $18.2_{(1.0)}$ | $16.7_{(0.9)}$ | $12.2_{(1.0)}$ | $16.4_{(1.3)}$ | $10.4_{(2.3)}$ | $9.5_{(0.5)}$ | $17.6_{(1.0)}$ |
| $p = 200$, $n = 400$ | NERCOME | $2.8_{(0.5)}$ | $2.5_{(0.4)}$ | $1.2_{(1.1)}$ | $0.9_{(1.2)}$ | $3.2_{(1.2)}$ | $1.2_{(0.2)}$ | $2.2_{(0.4)}$ |
| | CRC | $3.6_{(0.5)}$ | $3.3_{(0.5)}$ | $4.7_{(0.9)}$ | $4.5_{(1.1)}$ | $2.1_{(1.2)}$ | $1.5_{(0.3)}$ | $3.7_{(0.6)}$ |
| | NONLIN | $\mathbf{1.7_{(0.4)}}$ | $\mathbf{1.7_{(0.5)}}$ | $\mathbf{0.6_{(0.9)}}$ | $\mathbf{0.8_{(1.0)}}$ | $\mathbf{1.6_{(1.3)}}$ | $\mathbf{0.8_{(0.2)}}$ | $\mathbf{1.5_{(0.4)}}$ |
| | SCAD | $18.8_{(0.7)}$ | $17.0_{(0.7)}$ | $15.7_{(0.9)}$ | $19.4_{(1.2)}$ | $8.6_{(2.3)}$ | $9.6_{(0.4)}$ | $18.6_{(0.8)}$ |
| $p = 200$, $n = 800$ | NERCOME | $2.6_{(0.3)}$ | $2.5_{(0.3)}$ | $4.0_{(1.0)}$ | $3.1_{(0.9)}$ | $1.9_{(1.0)}$ | $1.1_{(0.2)}$ | $2.6_{(0.4)}$ |
| | CRC | $4.6_{(0.5)}$ | $4.9_{(0.5)}$ | $9.5_{(1.0)}$ | $8.1_{(1.0)}$ | $1.7_{(1.2)}$ | $2.1_{(0.2)}$ | $5.4_{(0.6)}$ |
| | NONLIN | $1.5_{(0.3)}$ | $1.5_{(0.3)}$ | $1.0_{(0.8)}$ | $1.0_{(0.8)}$ | $1.2_{(0.9)}$ | $0.7_{(0.2)}$ | $1.4_{(0.3)}$ |
| | SCAD | $\mathbf{6.6_{(1.4)}}$ | $\mathbf{5.6_{(1.4)}}$ | $\mathbf{5.5_{(1.6)}}$ | $\mathbf{5.7_{(1.6)}}$ | $\mathbf{6.0_{(2.8)}}$ | $\mathbf{3.6_{(0.7)}}$ | $\mathbf{6.1_{(1.4)}}$ |
| Profile (V) | | | | | | | | |
| $p = 100$, $n = 200$ | NERCOME | $7.2_{(2.4)}$ | $6.9_{(2.5)}$ | $\mathbf{0.5_{(2.6)}}$ | $\mathbf{0.3_{(3.2)}}$ | $8.0_{(3.4)}$ | $3.2_{(1.1)}$ | $5.5_{(1.9)}$ |
| | CRC | $\mathbf{5.7_{(2.2)}}$ | $\mathbf{5.2_{(1.9)}}$ | $5.1_{(3.2)}$ | $5.1_{(4.1)}$ | $\mathbf{4.7_{(3.1)}}$ | $\mathbf{2.4_{(1.1)}}$ | $\mathbf{5.4_{(2.4)}}$ |
| | NONLIN | $8.1_{(10.5)}$ | $6.4_{(3.4)}$ | $1.0_{(3.1)}$ | $2.9_{(7.1)}$ | $9.5_{(15.1)}$ | $3.7_{(3.5)}$ | $6.4_{(7.8)}$ |
| | SCAD | $39.0_{(7.6)}$ | $34.3_{(2.8)}$ | $25.4_{(2.4)}$ | $34.2_{(5.5)}$ | $33.2_{(15.6)}$ | $20.7_{(3.1)}$ | $36.7_{(5.7)}$ |
| $p = 100$, $n = 400$ | NERCOME | $\mathbf{5.0_{(1.4)}}$ | $4.7_{(1.3)}$ | $3.4_{(2.7)}$ | $2.8_{(2.6)}$ | $5.2_{(2.7)}$ | $\mathbf{2.2_{(0.6)}}$ | $4.3_{(1.1)}$ |
| | CRC | $6.2_{(1.1)}$ | $6.4_{(1.1)}$ | $10.6_{(2.6)}$ | $9.2_{(2.7)}$ | $\mathbf{3.8_{(2.6)}}$ | $2.7_{(0.5)}$ | $6.8_{(1.2)}$ |
| | NONLIN | $5.1_{(4.0)}$ | $\mathbf{4.4_{(1.5)}}$ | $\mathbf{1.0_{(2.0)}}$ | $\mathbf{2.1_{(2.9)}}$ | $5.4_{(10.7)}$ | $2.4_{(1.1)}$ | $\mathbf{4.2_{(2.6)}}$ |
| | SCAD | $11.7_{(6.8)}$ | $10.6_{(4.0)}$ | $3.3_{(4.0)}$ | $7.8_{(5.6)}$ | $14.9_{(16.3)}$ | $4.9_{(2.5)}$ | $10.5_{(5.4)}$ |
| $p = 100$, $n = 800$ | NERCOME | $4.3_{(0.8)}$ | $4.3_{(0.7)}$ | $7.0_{(2.3)}$ | $4.9_{(2.0)}$ | $3.6_{(2.6)}$ | $1.9_{(0.4)}$ | $4.3_{(0.7)}$ |
| | CRC | $7.1_{(1.0)}$ | $8.4_{(1.0)}$ | $17.4_{(2.6)}$ | $12.5_{(2.1)}$ | $4.1_{(2.7)}$ | $3.3_{(0.5)}$ | $8.4_{(1.0)}$ |
| | NONLIN | $4.4_{(4.2)}$ | $3.9_{(1.4)}$ | $1.5_{(1.7)}$ | $2.4_{(2.7)}$ | $4.0_{(9.4)}$ | $2.0_{(1.0)}$ | $3.8_{(2.7)}$ |
| | SCAD | $\mathbf{9.9_{(8.0)}}$ | $\mathbf{8.7_{(4.7)}}$ | $\mathbf{16.3_{(5.1)}}$ | $\mathbf{14.5_{(6.8)}}$ | $\mathbf{1.0_{(17.4)}}$ | $\mathbf{6.0_{(2.5)}}$ | $\mathbf{10.8_{(6.4)}}$ |
| $p = 200$, $n = 200$ | NERCOME | $4.3_{(1.4)}$ | $4.0_{(1.5)}$ | $\mathbf{0.9_{(1.8)}}$ | $\mathbf{1.3_{(2.2)}}$ | $5.7_{(2.4)}$ | $1.9_{(0.6)}$ | $2.8_{(0.9)}$ |
| | CRC | $\mathbf{3.2_{(0.8)}}$ | $\mathbf{2.8_{(0.7)}}$ | $1.9_{(1.8)}$ | $1.9_{(2.3)}$ | $\mathbf{3.3_{(1.8)}}$ | $\mathbf{1.4_{(0.3)}}$ | $\mathbf{2.7_{(0.8)}}$ |
| | *NONLIN | $5.0_{(6.8)}$ | $4.0_{(5.9)}$ | $1.0_{(8.7)}$ | $1.3_{(4.5)}$ | $9.3_{(16.6)}$ | $2.1_{(1.6)}$ | $3.7_{(4.7)}$ |
| | SCAD | $34.3_{(6.3)}$ | $29.1_{(2.2)}$ | $19.0_{(1.8)}$ | $27.5_{(4.1)}$ | $41.6_{(17.5)}$ | $17.4_{(2.0)}$ | $31.2_{(4.2)}$ |
| $p = 200$, $n = 400$ | NERCOME | $\mathbf{2.9_{(0.6)}}$ | $2.6_{(0.6)}$ | $1.5_{(1.8)}$ | $1.3_{(2.0)}$ | $3.0_{(1.8)}$ | $\mathbf{1.3_{(0.3)}}$ | $\mathbf{2.4_{(0.7)}}$ |
| | CRC | $3.5_{(0.6)}$ | $3.3_{(0.5)}$ | $4.7_{(1.4)}$ | $4.6_{(1.8)}$ | $\mathbf{1.9_{(1.8)}}$ | $1.5_{(0.3)}$ | $3.7_{(0.6)}$ |
| | NONLIN | $3.5_{(4.9)}$ | $\mathbf{2.5_{(1.1)}}$ | $\mathbf{0.0_{(1.4)}}$ | $\mathbf{0.9_{(2.6)}}$ | $6.3_{(14.6)}$ | $1.5_{(0.9)}$ | $2.5_{(2.7)}$ |
| | SCAD | $29.7_{(4.4)}$ | $25.6_{(1.5)}$ | $20.3_{(1.3)}$ | $27.2_{(2.7)}$ | $29.6_{(16.4)}$ | $15.2_{(1.2)}$ | $28.1_{(2.8)}$ |
| $p = 200$, $n = 800$ | NERCOME | $2.6_{(0.4)}$ | $2.5_{(0.3)}$ | $4.0_{(1.3)}$ | $3.1_{(1.4)}$ | $1.8_{(1.5)}$ | $\mathbf{1.1_{(0.2)}}$ | $2.6_{(0.4)}$ |
| | CRC | $4.6_{(0.5)}$ | $4.9_{(0.5)}$ | $9.4_{(1.3)}$ | $8.1_{(1.4)}$ | $\mathbf{1.6_{(1.6)}}$ | $2.0_{(0.2)}$ | $5.4_{(0.6)}$ |
| | NONLIN | $\mathbf{2.5_{(4.4)}}$ | $\mathbf{1.9_{(0.8)}}$ | $\mathbf{0.1_{(1.1)}}$ | $\mathbf{0.8_{(2.2)}}$ | $3.5_{(10.6)}$ | $1.1_{(0.7)}$ | $\mathbf{1.9_{(2.4)}}$ |
| | SCAD | $7.2_{(5.2)}$ | $5.8_{(1.9)}$ | $5.7_{(2.0)}$ | $8.0_{(3.3)}$ | $11.8_{(16.9)}$ | $2.7_{(1.2)}$ | $7.0_{(3.4)}$ |

Table 1: Mean efficiency loss (%) with respect to different loss functions for NERCOME, CRC grand average NONLIN and SCAD for profile (IV) and (V). The standard deviation is in bracket (%). $L_1$ =Frobenius loss, $L_2$ =inverse Stein's loss, $L_3$ =Frobenius loss of inverse, $L_4$ =Stein's loss, $L_5$ =Spectral norm of difference, $L_6$ =nuclear norm of difference, $L_7$ =Fréchet loss. Shaded cells mean the number is negative. Bolded cells indicate the minimum among all methods. For profile (V), when $p = n = 200$, NONLIN has two cases of singular covariance matrix, which has to be removed for mean and standard deviation calculations.

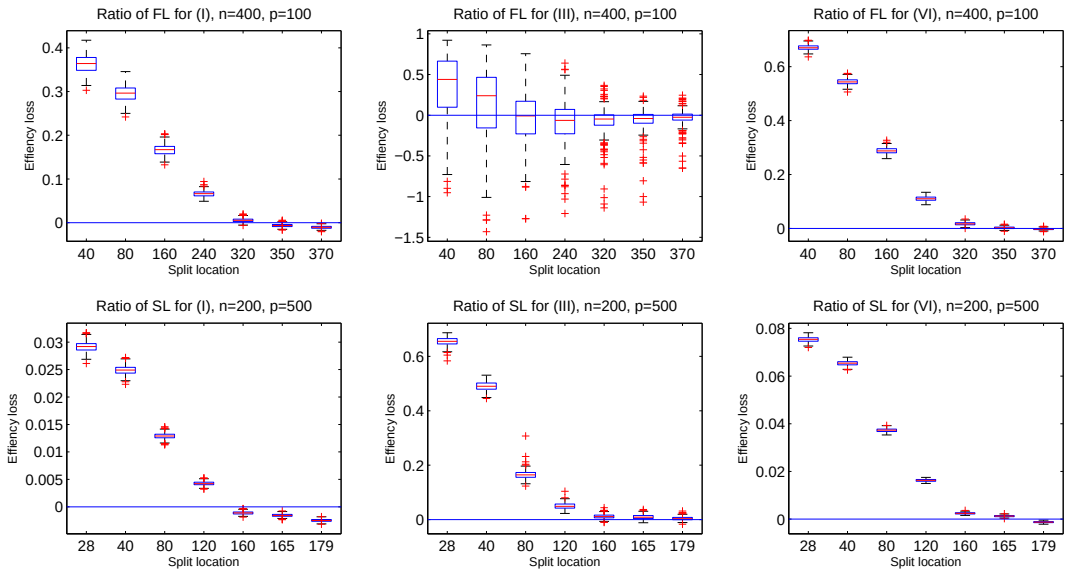


Figure 4: Boxplots of efficiency loss of $M^{-1}\sum_{i=1}^{M}\widehat{\mathbf{\Sigma}}_{Ideal,m}^{(i)}$ at split locations defined in (4.8). Upper row: With respect to the Frobenius loss, with $n = 400, p = 100$. Lower row: With respect to the inverse Stein's loss, with $n = 200, p = 500$. Left panel : For profile (I). Right panel: For profile (III).

because it is important for the ratio of loss $L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\text{Ideal},m,M})/L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\text{Ideal}})$ to be going to 1 almost surely (i.e., $EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\text{Ideal},m,M}) \overset{a.s.}{\to} 0$, which is what Figure 4 is showing) to ensure our estimator $\widehat{\mathbf{\Sigma}}_{m,M}$ to be almost surely asymptotically efficient. From the figure, it is clear that for the profiles (I) and (III), the split location $m$ should be close to $n$ for $EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\text{Ideal},m,M})$ to be close to 0. The efficiency loss has a wide variability for profile (III) with respect to the Frobenius loss (the upper right panel of Figure 4). It certainly achieves 0 on average for several split locations not close to $n$, so that the best $m$ is not always close to $n$ for factor model with fat-tailed error distribution.

For profile (I), with negative efficiency loss in almost all 500 simulation runs for the averaged ideal estimator $\widehat{\mathbf{\Sigma}}_{\text{Ideal},m,M}$ when split location is close to $n$ (see the left panels of Figure 4), it is clear that using the eigenmatrix $\mathbf{P}_1$ to form our estimator and averaging over different permutations of the data can be better than just using $\mathbf{P}$ from the full data set. Our estimator $\widehat{\mathbf{\Sigma}}_{m,M}$ in (4.6) follows the same spirit exactly.

## 2 Proof of theorems

We state an important Lemma first, which is in fact Lemma 2.7 of Bai and Silverstein (1998). Then, we present two more lemmas before presenting the proof of Theorem 1.

**Lemma S.1** *Let* $\mathbf{y} = (y_1, \ldots, y_p)^{\mathrm{T}}$ *be a complex random vector with independent and identically distributed (i.i.d.) entries satisfying* $E(y_1) = 0$, $E|y_1|^2 = 1$. *Let* $\mathbf{A}$ *be a given* $p \times p$ *matrix with possibly complex entries. Then for any* $q \geq 2$,

$$E|\mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{y} - tr(\mathbf{A})|^q \leq K_q \left( E^{q/2}|y_1|^4 \mathrm{tr}^{q/2}(\mathbf{A}\mathbf{A}^*) + E|y_1|^{2q}\mathrm{tr}(\mathbf{A}\mathbf{A}^*)^{q/2} \right),$$

*where* $K_q$ *is a constant depending only on* $q$, *and* $\mathbf{A}^* = \bar{\mathbf{A}}^{\mathrm{T}}$ *is the Hermitian of* $\mathbf{A}$.

**Lemma S.2** *For* $z \in \mathbb{C}^+$, *we have* $\left\|R_1(z)\right\|^2 \leq \frac{1}{\mathrm{Im}^2(z)}$.

*Proof of Lemma S.2.* Denote $\overline{z}$ the conjugate of $z$, and let $\widetilde{\mathbf{\Sigma}}_1 = \mathbf{P}_1\mathbf{D}_1\mathbf{P}_1^{\mathrm{T}}$. Then

$$\begin{aligned}
\left\|R_1(z)\right\|^2 &= \lambda_{\max}(R_1(z)\overline{R_1(z)}) \\
&= \lambda_{\max}(\mathbf{P}_1(\mathbf{D}_1 - z\mathbf{I}_p)^{-1}\mathbf{P}_1^{\mathrm{T}}\mathbf{P}_1(\mathbf{D}_1 - \overline{z}\mathbf{I}_p)^{-1}\mathbf{P}_1^{\mathrm{T}}) \\
&= \lambda_{\max}\left(\{[\mathbf{D}_1 - \mathrm{Re}(z)]^2 + \mathrm{Im}^2(z)\}^{-1}\right) \\
&\leq \frac{1}{\mathrm{Im}^2(z)}. \quad \square
\end{aligned}$$

**Lemma S.3** *Let* $\mathbf{R}$ *be Hermitian, that is, it is symmetric with complex entries. Then for a real square matrix* $\mathbf{B}$, *we have* $|\mathrm{tr}(\mathbf{B}\mathbf{R})| \leq |\lambda_{\max}(\mathbf{B})||\mathrm{tr}(\mathbf{R})|$. *Also,* $|\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{R}\mathbf{A})| \leq 2^{1/2}\|\mathbf{R}\|\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{A})$ *for a real matrix* $\mathbf{A}$.

*Proof of Lemma S.3.* Since $\mathbf{R}$ is Hermitian, it can be decomposed as $\mathbf{R} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$, where $\mathbf{U}$ is orthogonal and $\mathbf{D}$ is diagonal with complex entries. Writing $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)$ and $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_p)$, then

$$\begin{aligned}
|\mathrm{tr}(\mathbf{B}\mathbf{R})| = |\mathrm{tr}(\mathbf{U}^{\mathrm{T}}\mathbf{B}\mathbf{U}\mathbf{D})| &= \left|\sum_{i=1}^p d_i \mathbf{u}_i^{\mathrm{T}}\mathbf{B}\mathbf{u}_i\right| \\
&\leq |\lambda_{\max}(\mathbf{B})| \cdot \left|\sum_{i=1}^p d_i\right| = |\lambda_{\max}(\mathbf{B})||\mathrm{tr}(\mathbf{R})|.
\end{aligned}$$

Finally, writing $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_r)$,

$$\begin{aligned}
|\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{R}\mathbf{A})| = \left|\sum_{i=1}^r \mathbf{a}_i^{\mathrm{T}}\mathbf{R}\mathbf{a}_i\right| &= \left( \left[\sum_{i=1}^r \mathbf{a}_i^{\mathrm{T}}\mathrm{Re}(\mathbf{R})\mathbf{a}_i\right]^2 + \left[\sum_{i=1}^r \mathbf{a}_i^{\mathrm{T}}\mathrm{Im}(\mathbf{R})\mathbf{a}_i\right]^2 \right)^{1/2} \\
&\leq \left(\lambda_{\max}^2[\mathrm{Re}(\mathbf{R})] + \lambda_{\max}^2[\mathrm{Im}(\mathbf{R})]\right)^{1/2} \mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) \\
&\leq \left(2\lambda_{\max}[\mathrm{Re}^2(\mathbf{R}) + \mathrm{Im}^2(\mathbf{R})]\right)^{1/2} \mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) \\
&= 2^{1/2}\|\mathbf{R}\|\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{A}). \quad \square
\end{aligned}$$

*Proof of Theorem 1.* We prove part (i) first. Consider $\mathbf{\Psi}_m^{(1)}(z) - \mathbf{\Psi}^{(1)}(z) = I_1 + I_2$, where

$$I_1 = \mathbf{\Psi}_m^{(1)}(z) - \frac{1}{p}\mathrm{tr}\left(R_1(z)\mathbf{\Sigma}_p\right), \quad I_2 = \frac{1}{p}\mathrm{tr}\left(R_1(z)\mathbf{\Sigma}_p\right) - \mathbf{\Psi}^{(1)}(z),$$

with $R_1(z) = (\widetilde{\mathbf{\Sigma}}_1 - z\mathbf{I}_p)^{-1}$. A direct application of Lemma 2 of Ledoit and Péché (2011) implies that $p^{-1}\mathrm{tr}(R_1(z)\mathbf{\Sigma}_p)$ converges a.s. to $\mathbf{\Psi}^{(1)}(z)$, and hence $I_2 \overset{a.s.}{\to} 0$. It remains to show that $I_1 \overset{a.s.}{\to} 0$ as well.

To this end, write $\mathbf{Y}_2 = (\mathbf{y}_{21}, \ldots, \mathbf{y}_{2n_2})$, so that $\widetilde{\mathbf{\Sigma}}_2 = n_2^{-1}\sum_{k=1}^{n_2}\mathbf{y}_{2k}\mathbf{y}_{2k}^{\mathrm{T}}$. Also, write $\mathbf{y}_{2k} = \mathbf{\Sigma}_p^{1/2}\mathbf{z}_{2k}$ for $k = 1, \ldots, n_2$ with $\mathbf{z}_{2k}$ being one of the $\mathbf{z}_i$'s in assumption (A1). Using the independence of the $\mathbf{z}_{2k}$'s and $\mathbf{Y}_1$, we have for $k = 1, \ldots, n_2$, $2 \le q \le 6$ and $z \in \mathbb{C}^+$,

$$\begin{aligned}
&E\left\{\left|\mathbf{z}_{2k}^{\mathrm{T}}\mathbf{\Sigma}_p^{1/2}R_1(z)\mathbf{\Sigma}_p^{1/2}\mathbf{z}_{2k} - \mathrm{tr}(\mathbf{\Sigma}_p^{1/2}R_1(z)\mathbf{\Sigma}_p^{1/2})\right|^q \Big| \mathbf{Y}_1\right\} \\
&\le K_q(E^{q/2}|z_{2k}|^4\mathrm{tr}^{q/2}(\mathbf{\Sigma}_p R_1(z)\mathbf{\Sigma}_p\overline{R_1(z)}) \\
&\quad + E|z_{2k}|^{2q}\mathrm{tr}(\mathbf{\Sigma}_p^{1/2}R_1(z)\mathbf{\Sigma}_p\overline{R_1(z)}\mathbf{\Sigma}_p^{1/2})^{q/2}) \\
&\le K_q(E^{q/2}|z_{2k}|^4\left\|\mathbf{\Sigma}_p\right\|^q p^{q/2}\mathrm{Im}^{-q}(z) \\
&\quad + E|z_{2k}|^{2q}p\lambda_{\max}^{q/2}(\mathbf{\Sigma}_p^{1/2}R_1(z)\mathbf{\Sigma}_p\overline{R_1(z)}\mathbf{\Sigma}_p^{1/2})) \\
&\le K_p\mathrm{Im}^{-q}(z)p^{q/2}\left\|\mathbf{\Sigma}_p\right\|^q(E^{q/2}|z_{2k}|^4 + p^{1-q/2}E|z_{2k}|^{2q}) = O(p^q),
\end{aligned}$$

where the second line uses Lemma S.1, which is applicable by assumption (A1) on $\mathbf{z}_{2k}$, and the second last line is by Lemma S.3. The last line uses the assumptions in (A1). We have

$$\begin{aligned}
E(|I_1|^6|\mathbf{Y}_1) &= E\left(\left|\frac{1}{pn_2}\mathrm{tr}\left(R_1(z)\sum_{k=1}^{n_2}\mathbf{y}_{2k}\mathbf{y}_{2k}^{\mathrm{T}}\right) - \frac{1}{p}\mathrm{tr}(R_1(z)\mathbf{\Sigma}_p)\right|^6 \Big| \mathbf{Y}_1\right) \\
&= E\left(\left|\frac{1}{pn_2}\sum_{k=1}^{n_2}(\mathbf{z}_{2k}^{\mathrm{T}}\mathbf{\Sigma}_p^{1/2}R_1(z)\mathbf{\Sigma}_p^{1/2}\mathbf{z}_{2k} - \mathrm{tr}(R_1(z)\mathbf{\Sigma}_p))\right|^6 \Big| \mathbf{Y}_1\right).
\end{aligned}$$

Define $g_k = \mathbf{z}_{2k}^{\mathrm{T}}\mathbf{\Sigma}_p^{1/2}R_1(z)\mathbf{\Sigma}_p^{1/2}\mathbf{z}_{2k} - \mathrm{tr}(R_1(z)\mathbf{\Sigma}_p)$. Then $E(g_i|\mathbf{Y}_1) = 0$, and $g_i|\mathbf{Y}_1$ is independent of $g_j|\mathbf{Y}_1$ for $i \ne j$. We then have the expansion

$$\begin{aligned}
E(|I_1|^6) &= \frac{1}{p^6n_2^6}\sum_{i_1,\ldots,i_6=1}^{n_2}E(E(g_{i_1}\bar{g}_{i_2}g_{i_3}\bar{g}_{i_4}g_{i_5}\bar{g}_{i_6}|\mathbf{Y}_1)) \\
&\le \frac{1}{p^6n_2^6}E\Bigg(\sum_{i=1}^{n_2}E(|g_i|^6|\mathbf{Y}_1) + \sum_{i\ne j}[E(|g_i|^2|\mathbf{Y}_1)E(|g_j|^4|\mathbf{Y}_1) \\
&\quad + E(|g_i|^3|\mathbf{Y}_1)E(|g_j|^3|\mathbf{Y}_1)] \\
&\quad + \sum_{i\ne j\ne k}E(|g_i|^2|\mathbf{Y}_1)E(|g_j|^2|\mathbf{Y}_1)E(|g_k|^2|\mathbf{Y}_1)\Bigg) \qquad \text{(S.1)} \\
&= O(p^{-6}n_2^{-6}\cdot[n_2 + 2n_2(n_2-1) + n_2(n_2-1)(n_2-2)]p^6) = O(n_2^{-3}),
\end{aligned}$$

where the last line used $E(|g_i|^q|\mathbf{Y}_1) = O(p^q)$ proved before. Since $\sum_{n\geq 1} n_2^{-3} < \infty$ by assumption, we can apply the Borel-Cantelli lemma to conclude that $I_1 \overset{a.s.}{\to} 0$. This completes the proof of part (i).

Note also that since $I_1 \overset{a.s.}{\to} 0$, we have $\mathbf{\Psi}_m^{(1)}(z) - p^{-1}\mathrm{tr}(R_1(z)\mathbf{\Sigma}_p) \overset{a.s.}{\to} 0$. Using the inversion formula (2.3), it is easy to show that the inverse Stieltjes transform for $\mathbf{\Psi}_m^{(1)}(z)$ is $\mathbf{\Phi}_m^{(1)}(x)$ in (2.13) on all points of continuity of $\mathbf{\Phi}_m^{(1)}(x)$, and that for $p^{-1}\mathrm{tr}(R_1(z)\mathbf{\Sigma}_p)$ is $p^{-1}\sum_{i=1}^p \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}\mathbf{1}_{\{\lambda_{1i}\leq x\}}$ on all points of continuity of the function. Hence, using equation (2.5) of Silverstein and Bai (1995), we conclude that

$$\mathbf{\Phi}_m^{(1)}(x) - \frac{1}{p}\sum_{i=1}^p \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}\mathbf{1}_{\{\lambda_{1i}\leq x\}} \overset{a.s.}{\to} 0$$

as $n_1, p \to \infty$ with $p/n_1 \to c_1 > 0$, which certainly include the case $c_1 = 1$. This proves the last part of the theorem.

The proof for part (ii) of the theorem can be found in the proof of Theorem 4 of Ledoit and Péché (2011), from equations (45) to (50) by replacing $\gamma$ there with $1/c_1$ and $F$ with $F_1$. Part (iii) is just an application of Theorem 4 of Ledoit and Péché (2011). This completes the proof of the theorem. □

*Proof of Theorem 3.* Let $\mathbf{Y}_2 = (\mathbf{y}_{21}, \ldots, \mathbf{y}_{2n_2})$ with $\mathbf{y}_{2k} = \mathbf{A}\mathbf{x}_{2k} + \boldsymbol{\epsilon}_{2k}$, $k = 1, \ldots, n_2$. Let $R_1(z) = (\widetilde{\mathbf{\Sigma}}_1 - z\mathbf{I}_p)^{-1}$. Then with $\mathbf{\Sigma}_p = \mathbf{A}\mathbf{\Sigma}_x\mathbf{A} + \mathbf{\Sigma}_\epsilon$ as in (3.18), we have

$$\begin{aligned}
&\frac{1}{p}\mathrm{tr}(R_1(z)\widetilde{\mathbf{\Sigma}}_2) - \frac{1}{p}\mathrm{tr}(R_1(z)\mathbf{\Sigma}_p) = D_1 + D_2 + D_3, \text{ where}\\
&D_1 = \frac{1}{pn_2}\sum_{k=1}^{n_2}\left\{\mathbf{x}_{2k}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}R_1(z)\mathbf{A}\mathbf{x}_{2k} - \mathrm{tr}(R_1(z)\mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^{\mathrm{T}})\right\},\\
&D_2 = \frac{2}{pn_2}\sum_{k=1}^{n_2}\mathbf{x}_{2k}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}R_1(z)\boldsymbol{\epsilon}_{2k},\\
&D_3 = \frac{1}{pn_2}\sum_{k=1}^{n_2}\{\boldsymbol{\epsilon}_{2k}^{\mathrm{T}}R_1(z)\boldsymbol{\epsilon}_{2k} - \mathrm{tr}(R_1(z)\mathbf{\Sigma}_\epsilon)\}.
\end{aligned}$$

Consider $D_1$ first. Using assumption (F1), and defining

$$g_i = \mathbf{x}_{2k}^{*\mathrm{T}}\mathbf{\Sigma}_x^{1/2}\mathbf{A}^{\mathrm{T}}R_1(z)\mathbf{A}\mathbf{\Sigma}_x^{1/2}\mathbf{x}_{2k}^* - \mathrm{tr}(R_1(z)\mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^{\mathrm{T}}),$$

for $2 \leq q \leq 6$ and $z \in \mathbb{C}^+$, we have by Lemma S.1,

$$\begin{aligned}
E(|g_k|^q|\mathbf{Y}_1) &\leq K_q(E^{q/2}|x_{2k}^*|^4\mathrm{tr}^{q/2}(\mathbf{\Sigma}_x^{1/2}\mathbf{A}^{\mathrm{T}}R_1(z)\mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^{\mathrm{T}}\overline{R_1(z)}\mathbf{A}\mathbf{\Sigma}_x^{1/2})\\
&\quad + E|x_{2k}^*|^{2q}\mathrm{tr}(\mathbf{\Sigma}_x^{1/2}\mathbf{A}^{\mathrm{T}}R_1(z)\mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^{\mathrm{T}}\overline{R_1(z)}\mathbf{A}\mathbf{\Sigma}_x^{1/2})^{q/2})\\
&\leq K_q\left\|\mathbf{\Sigma}_x^{1/2}\mathbf{A}^{\mathrm{T}}R_1(z)\mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^{\mathrm{T}}\overline{R_1(z)}\mathbf{A}\mathbf{\Sigma}_x^{1/2}\right\|^{q/2}\\
&\quad \cdot (E^{q/2}|x_{2k}^*|^4 r^{q/2} + rE|x_{2k}^*|^{2q})\\
&\leq K_q\left\|\mathbf{\Sigma}_x\right\|^q\left\|\mathbf{A}\right\|^{2q}\mathrm{Im}^{-q}(z)(E^{q/2}|x_{2k}^*|^4 r^{q/2} + rE|x_{2k}^*|^{2q}) = O(p^q),
\end{aligned}$$

where we have used $\|\mathbf{A}\| = O(p^{1/2})$ implied in assumption (F2). Using the fact that $E(g_k|\mathbf{Y}_1) = 0$ and $g_i|\mathbf{Y}_1$ is independent of $g_j|\mathbf{Y}_1$ for $i \neq j$ by assumption (F1), similar to the expansion in (S.1), we have

$$E(|D_1|^6) = \frac{1}{p^6 n_2^6} \sum_{i_1 \cdots i_6 = 1}^{n_2} E(E(g_{i_1}\bar{g}_{i_2} g_{i_3}\bar{g}_{i_4} g_{i_5}\bar{g}_{i_6}|\mathbf{Y}_1)) = O(n_2^{-3}),$$

so that $D_1 \overset{a.s.}{\to} 0$ by the Borel-Cantelli lemma, as $\sum_{n\geq 1} n_2^{-3} < \infty$ by assumption.

Now consider $D_2$. Define $g_k = \mathbf{x}_{2k}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} R_1(z) \boldsymbol{\epsilon}_{2k}$. We have for $0 < q \leq 12$ and $z \in \mathbb{C}^+$,

$$E(|g_k|^q|\mathbf{Y}_1) \leq E\|\mathbf{x}_{2k}^*\|^q \|\boldsymbol{\Sigma}_x\|^{q/2} \|\mathbf{A}\|^q E\|\boldsymbol{\xi}_{2k}\|^q \|\boldsymbol{\Sigma}_\epsilon\|^{q/2} \mathrm{Im}^{-q}(z) = O(p^q),$$

by assumptions (F1) and (F2), and lemma S.2. Hence, using the fact that $E(g_k|\mathbf{Y}_1) = 0$ and $g_i|\mathbf{Y}_1$ is independent of $g_j|\mathbf{Y}_1$ for $i \neq j$, an expansion similar to (S.1) leads again to

$$E(|D_2|^6) = O(n_2^{-3}),$$

so that by the Borel-Cantelli lemma, we have $D_2 \overset{a.s.}{\to} 0$.

Finally, if we define $g_k = \boldsymbol{\xi}_{2k}^{\mathrm{T}} \boldsymbol{\Sigma}_\epsilon^{1/2} R_1(z) \boldsymbol{\Sigma}_\epsilon^{1/2} \boldsymbol{\xi}_{2k} - \mathrm{tr}(R_1(z)\boldsymbol{\Sigma}_\epsilon)$, then following exactly the same arguments as in the proof of Theorem 1, we can conclude

$$E(|D_3|^6) = O(n_2^{-3}),$$

so that through the Borel-Cantelli lemma we have $D_3 \overset{a.s.}{\to} 0$.

Hence, we have proved

$$\frac{1}{p}\mathrm{tr}[R_1(z)\widetilde{\boldsymbol{\Sigma}}_2] - \frac{1}{p}\mathrm{tr}[R_1(z)\boldsymbol{\Sigma}_p] \overset{a.s.}{\to} 0.$$

Then, using equation (2.5) of Silverstein and Bai (1995), we conclude the almost sure convergence of the inverse Stieltjes transform on all points of continuity $x$, that is,

$$\frac{1}{p}\sum_{i=1}^{p} \mathbf{p}_{1i}^{\mathrm{T}} \widetilde{\boldsymbol{\Sigma}}_2 \mathbf{p}_{1i} \mathbf{1}_{\{\lambda_{1i} \leq x\}} - \frac{1}{p}\sum_{i=1}^{p} \mathbf{p}_{1i}^{\mathrm{T}} \boldsymbol{\Sigma}_p \mathbf{p}_{1i} \mathbf{1}_{\{\lambda_{1i} \leq x\}} \overset{a.s.}{\to} 0. \quad \square$$

*Proof of Lemma 1.* We first assume (A1)' and (A2)'. For $j = 1, \ldots, n_2$ and $i = 1, \ldots, p$, define

$$g_{ij} = \frac{\mathbf{z}_{2j}^{\mathrm{T}} \boldsymbol{\Sigma}_p^{1/2} \mathbf{p}_{1i} \mathbf{p}_{1i}^{\mathrm{T}} \boldsymbol{\Sigma}_p^{1/2} \mathbf{z}_{2j} - \mathrm{tr}(\boldsymbol{\Sigma}_p^{1/2} \mathbf{p}_{1i} \mathbf{p}_{1i}^{\mathrm{T}} \boldsymbol{\Sigma}_p^{1/2})}{\mathbf{p}_{1i}^{\mathrm{T}} \boldsymbol{\Sigma}_p \mathbf{p}_{1i}}.$$

Then for $2 \leq q \leq 10$, using Lemma S.1 and assumption (A1)',

$$\begin{aligned} E(|g_{ij}|^q|\mathbf{Y}_1) &\leq (\mathbf{p}_{1i}^{\mathrm{T}} \boldsymbol{\Sigma}_p \mathbf{p}_{1i})^{-q} K_q \Big( E^{q/2}|z_{2j}|^4 \mathrm{tr}^q(\boldsymbol{\Sigma}_p^{1/2} \mathbf{p}_{1i} \mathbf{p}_{1i}^{\mathrm{T}} \boldsymbol{\Sigma}_p^{1/2}) \\ &\quad + E|z_{2j}|^{2q} \mathrm{tr}(\boldsymbol{\Sigma}_p^{1/2} \mathbf{p}_{1i} \mathbf{p}_{1i}^{\mathrm{T}} \boldsymbol{\Sigma}_p^{1/2})^q \Big) \\ &= K_q(E^{q/2}|z_{2j}|^4 + E|z_{2j}|^{2q}) = O(1). \end{aligned}$$

At the same time, $E(g_{ij}|\mathbf{Y}_1) = 0$ and $g_{i_1j_1}|\mathbf{Y}_1$ is independent of $g_{i_2j_2}|\mathbf{Y}_1$ whenever $j_1 \neq j_2$. Hence,

$$
\begin{aligned}
&E\left\{\max_{1\le i\le p}\left|\frac{\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}}\right|\right\}^{10} = E\left\{\max_{1\le i\le p}\left|\frac{1}{n_2}\sum_{j=1}^{n_2} g_{ij}\right|^{10}\right\} \\
&\le \sum_{i=1}^{p}\frac{1}{n_2^{10}}\sum_{j_1\cdots j_{10}=1}^{n_2} E(E\{g_{ij_1}\cdots g_{ij_{10}}|\mathbf{Y}_1\}) \\
&\le E\Bigg\{\sum_{i=1}^{p}\frac{1}{n_2^{10}}\Bigg(\sum_{j=1}^{n_2}E(g_{ij}^{10}|\mathbf{Y}_1) + \sum_{\substack{q_1,q_2\neq 1\\ q_1+q_2=10}}\sum_{j_1\neq j_2}E(g_{ij_1}^{q_1}|\mathbf{Y}_1)E(g_{ij_2}^{q_2}|\mathbf{Y}_1) \\
&+ \sum_{\substack{q_1,q_2,q_3\neq 1\\ q_1+q_2+q_3=10}}\sum_{j_1\neq j_2\neq j_3}\prod_{k=1}^{3}E(g_{ij_k}^{q_k}|\mathbf{Y}_1) + \sum_{\substack{q_1,\ldots,q_4\neq 1\\ q_1+\cdots+q_4=10}}\sum_{j_1\neq j_2\neq j_3\neq j_4}\prod_{k=1}^{4}E(g_{ij_k}^{q_k}|\mathbf{Y}_1) \\
&+ \sum_{j_1\neq\cdots\neq j_5}\prod_{k=1}^{5}E(g_{ij_k}^{2}|\mathbf{Y}_1)\Bigg)\Bigg\} \\
&= O(pn_2^{-10}[n_2 + 4\mathrm{P}_2^{n_2} + 4\mathrm{P}_3^{n_2} + 2\mathrm{P}_4^{n_2} + \mathrm{P}_5^{n_2}]) \\
&= O(pn_2^{-5}),
\end{aligned}
\tag{S.2}
$$

where $\mathrm{P}_r^{n_2} = n_2(n_2-1)\cdots(n_2-r+1)$. Since $\sum_{n\ge 1} pn_2^{-5} < \infty$ by assumption, the Borel-Cantelli lemma completes the proof of this part.

Now consider data from a factor model, with assumption (F1)' satisfied. We can decompose

$$
\begin{aligned}
&\frac{\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}} = D_1 + D_2 + D_3, \quad \text{where} \\
&D_1 = \frac{1}{n_2}\sum_{j=1}^{n_2} g_{ij}, \quad g_{ij} = \frac{\mathbf{x}_{2j}^{*\mathrm{T}}\mathbf{\Sigma}_x^{1/2}\mathbf{A}^{\mathrm{T}}\mathbf{p}_{1i}\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{A}\mathbf{\Sigma}_x^{1/2}\mathbf{x}_{2j}^{*} - \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^{\mathrm{T}}\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}}, \\
&D_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} d_{ij}, \quad d_{ij} = \frac{\boldsymbol{\xi}_{2j}^{\mathrm{T}}\mathbf{\Sigma}_\epsilon^{1/2}\mathbf{p}_{1i}\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_\epsilon^{1/2}\boldsymbol{\xi}_{2j} - \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_\epsilon\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}}, \\
&D_3 = \frac{2}{n_2}\sum_{j=1}^{n_2} h_{ij}, \quad h_{ij} = \frac{\mathbf{x}_{2j}^{*\mathrm{T}}\mathbf{\Sigma}_x^{1/2}\mathbf{A}^{\mathrm{T}}\mathbf{p}_{1i}\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_\epsilon^{1/2}\boldsymbol{\xi}_{2j}}{\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}}.
\end{aligned}
$$

Using Lemma S.1, for $2 \le q \le 10$,

$$
\begin{aligned}
E(|g_{ij}|^q|\mathbf{Y}_1) &\le (\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i})^{-q}K_q(\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{A}\mathbf{\Sigma}_x\mathbf{A}^{\mathrm{T}}\mathbf{p}_{1i})^q(E^{q/2}|x_{2j}^{*}|^4 + E|x_{2j}^{*}|^{2q}) \\
&\le K_q(E^{q/2}|x_{2j}^{*}|^4 + E|x_{2j}^{*}|^{2q}) = O(1); \\
E(|d_{ij}|^q|\mathbf{Y}_1) &\le (\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i})^{-q}K_q(\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_\epsilon\mathbf{p}_{1i})^q(E^{q/2}|\xi_{2j}|^4 + E|\xi_{2j}|^{2q}) \\
&\le K_q(E^{q/2}|\xi_{2j}|^4 + E|\xi_{2j}|^{2q}) = O(1).
\end{aligned}
$$

For $4 \leq q \leq 10$, using the $c_r$-inequality,

$$E(|h_{ij}|^q|\mathbf{Y}_1) = E\left(\left(g_{ij} + \frac{\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^{\mathrm{T}}\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}}\right)^{q/2}\Big|\mathbf{Y}_1\right) \cdot E\left(\left(d_{ij} + \frac{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}}\right)^{q/2}\Big|\mathbf{Y}_1\right), \text{ with}$$

$$E\left(\left(g_{ij} + \frac{\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^{\mathrm{T}}\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}}\right)^{q/2}\Big|\mathbf{Y}_1\right) \leq C_{q/2}(E(|g_{ij}|^{q/2}|\mathbf{Y}_1) + 1) = O(1),$$

$$E\left(\left(\frac{\boldsymbol{\xi}_{2j}^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon^{1/2}\mathbf{p}_{1i}\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon^{1/2}\boldsymbol{\xi}_{2j}}{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}}\right)^{q/2}\Big|\mathbf{Y}_1\right) \leq C_{q/2}(E(|d_{ij}|^{q/2}|\mathbf{Y}_1) + 1) = O(1),$$

where $C_{q/2}$ is a constant. Hence, $E(|h_{ij}|^q|\mathbf{Y}_1) = O(1)$ for $q \geq 4$. For $2 \leq q < 4$, it is easy to see that $E(|h_{ij}|^q|\mathbf{Y}_1) \leq E^{1/2}(|h_{ij}|^{2q}|\mathbf{Y}_1) = O(1)$, and hence $E(|h_{ij}|^q|\mathbf{Y}_1) = O(1)$ for $2 \leq q \leq 10$. Then since $E(g_{ij}|\mathbf{Y}_1) = 0$ and $g_{i_1j_1}|\mathbf{Y}_1$ is independent of $g_{i_2j_2}|\mathbf{Y}_1$ whenever $j_1 \neq j_2$, an expansion exactly like (S.2) will show that

$$E\left(\left|\max_{1\leq i\leq p}|D_1|^{10}\right|\right) = O(pn_2^{-5}),$$

showing that $\max_{1\leq i\leq p}|D_1| \overset{a.s.}{\to} 0$, since $\sum_{n\geq 1} pn_2^{-5} < \infty$ by assumption. Using the same expansion (S.2) and the arguments above, we can also show that $\max_{1\leq i\leq p}|D_2|$, $\max_{1\leq i\leq p}|D_3| \overset{a.s.}{\to} 0$. This completes the proof of the lemma. □

Before proving Theorem 5, we state and prove a lemma first.

**Lemma S.4** *Let assumptions (A1) to (A4) be satisfied. Then assuming* $\boldsymbol{\Sigma}_p \neq \sigma^2\mathbf{I}$,

$$\frac{\left\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{Ideal},m} - \boldsymbol{\Sigma}_p\right\|_F^2}{\left\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{Ideal}} - \boldsymbol{\Sigma}_p\right\|_F^2} \overset{a.s.}{\to} 1, \quad \frac{\log\det(\widehat{\boldsymbol{\Sigma}}_{\mathrm{Ideal},m}\boldsymbol{\Sigma}_p^{-1})}{\log\det(\widehat{\boldsymbol{\Sigma}}_{\mathrm{Ideal}}\boldsymbol{\Sigma}_p^{-1})} \overset{a.s.}{\to} 1.$$

*Proof of Lemma S.4.* Consider the Frobenius loss first. For $\lambda_i$ the $i$th largest eigenvalue of $\mathbf{S}_n$ with corresponding eigenvector $\mathbf{p}_i$,

$$\begin{aligned} p^{-1}\left\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{Ideal}} - \boldsymbol{\Sigma}_p\right\|_F^2 &= p^{-1}\mathrm{tr}(\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{P}) - \mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{P})^2 \\ &= p^{-1}\sum_{i=1}^{n}(\mathbf{p}_i^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_i)^2 - 2p^{-1}\mathrm{tr}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{P})) \\ &\quad + p^{-1}\mathrm{tr}(\boldsymbol{\Sigma}_p^2) \\ &= p^{-1}\mathrm{tr}(\boldsymbol{\Sigma}_p^2) - p^{-1}\sum_{i=1}^{p}(\mathbf{p}_i^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_i)^2, \end{aligned}$$

so that similarly, $p^{-1}\left\|\widehat{\mathbf{\Sigma}}_{\text{Ideal},m} - \mathbf{\Sigma}_p\right\|_F^2 = p^{-1}\text{tr}(\mathbf{\Sigma}_p^2) - p^{-1}\sum_{i=1}^p(\mathbf{p}_{1i}^{\text{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i})^2$. But with assumption (A3), we have

$$p^{-1}\text{tr}(\mathbf{\Sigma}_p^2) = p^{-1}\sum_{i=1}^p \tau_{n,i}^2 \overset{a.s.}{\to} \int \tau^2 dH(\tau).$$

Moreover, with the convergence result in Theorem 4 of Ledoit and Péché (2011) indicating that (see equation (2.7) as well) for all $x \in \mathbb{R}$, we have

$$\Delta_p(x) = p^{-1}\sum_{i=1}^p \mathbf{p}_i^{\text{T}}\mathbf{\Sigma}_p\mathbf{p}_i\mathbf{1}_{\{\lambda_i \le x\}} \overset{a.s.}{\to} \Delta(x) = \int_{-\infty}^x \delta(\lambda)dF(\lambda),$$

where $F(\cdot)$ is such that

$$F_p(\lambda) = p^{-1}\sum_{i=1}^p \mathbf{1}_{\{\lambda_i \le \lambda\}} \overset{a.s.}{\to} F(\lambda),$$

we have for any continuous function $g(\cdot)$ over the positive real line,

$$p^{-1}\sum_{i=1}^p g(\mathbf{p}_i^{\text{T}}\mathbf{\Sigma}_p\mathbf{p}_i)\mathbf{1}_{\{\lambda_i \le x\}} \overset{a.s.}{\to} \int_{-\infty}^x g(\delta(\lambda))dF(\lambda). \tag{S.3}$$

With this, it is easy to see that then

$$p^{-1}\left\|\widehat{\mathbf{\Sigma}}_{\text{Ideal}} - \mathbf{\Sigma}_p\right\|_F^2 \overset{a.s.}{\to} \int \tau^2 dH(\tau) - \int \delta^2(\lambda)dF(\lambda),$$

which is nonzero in general, except when $\mathbf{\Sigma}_p = \sigma^2\mathbf{I}$. At the same time, we have (see equation (2.9) and the descriptions therein as well)

$$p^{-1}\sum_{i=1}^p \mathbf{p}_{1i}^{\text{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i}\mathbf{1}_{\{\lambda_{1i} \le x\}} \overset{a.s.}{\to} \int_{-\infty}^x \delta_1(\lambda)dF_1(\lambda),$$

where $F_1(\cdot)$ is such that

$$F_{1p}(\lambda) = p^{-1}\sum_{i=1}^p \mathbf{1}_{\{\lambda_{1i} \le \lambda\}} \overset{a.s.}{\to} F_1(\lambda).$$

But since $m/n = n_1/n \to 1$, we have $p/n_1, p/n$ both going to the same limit $c_1 = c > 0$. Theorem 4.1 of Bai and Silverstein (2010) tells us then both $F_p$ and $F_{1p}$ converges to the same limit almost surely under assumptions (A1) to (A4). Hence $F = F_1$ almost surely,implying $\delta_1(\cdot) = \delta(\cdot)$ almost surely (See Remark 1 in section 2.2 as well). This immediately implies that, similar to (S.3),

$$p^{-1}\sum_{i=1}^p(\mathbf{p}_{1i}^{\text{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i})^2\mathbf{1}_{\{\lambda_{1i} \le x\}} \overset{a.s.}{\to} \int_{-\infty}^x \delta_1^2(\lambda)dF_1(\lambda) = \int_{-\infty}^x \delta^2(\lambda)dF(\lambda),$$

and hence

$$p^{-1}\left\|\widehat{\mathbf{\Sigma}}_{\text{Ideal},m} - \mathbf{\Sigma}_p\right\|_F^2 \overset{a.s.}{\to} \int \tau^2 dH(\tau) - \int \delta_1^2(\lambda) dF_1(\lambda)$$
$$= \int \tau^2 dH(\tau) - \int \delta^2(\lambda) dF(\lambda).$$

We immediately have

$$\frac{\left\|\widehat{\mathbf{\Sigma}}_{\text{Ideal},m} - \mathbf{\Sigma}_p\right\|_F^2}{\left\|\widehat{\mathbf{\Sigma}}_{\text{Ideal}} - \mathbf{\Sigma}_p\right\|_F^2} \overset{a.s.}{\to} 1,$$

which completes the proof for the Frobenius loss.

For the remaining part, note that it is not difficult to show the inverse Stein's loss

$$p^{-1} SL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\text{Ideal}}) = p^{-1} \log\det(\widehat{\mathbf{\Sigma}}_{\text{Ideal}} \mathbf{\Sigma}_p^{-1})$$
$$= p^{-1} \sum_{i=1}^{p} \log(\mathbf{p}_i^{\mathrm{T}} \mathbf{\Sigma}_p \mathbf{p}_i) - p^{-1} \sum_{i=1}^{p} \log(\tau_{n,i}).$$

Similar arguments as before show that

$$p^{-1} \sum_{i=1}^{p} \log(\mathbf{p}_i^{\mathrm{T}} \mathbf{\Sigma}_p \mathbf{p}_i) \overset{a.s.}{\to} \int \log(\delta(\lambda)) dF(\lambda),$$
$$p^{-1} \sum_{i=1}^{p} \log(\tau_{n,i}) \overset{a.s.}{\to} \int \log(\tau) dH(\tau).$$

Hence

$$p^{-1} SL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\text{Ideal}}) = p^{-1} \log\det(\widehat{\mathbf{\Sigma}}_{\text{Ideal}} \mathbf{\Sigma}_p^{-1})$$
$$\overset{a.s.}{\to} \int \log(\delta(\lambda)) dF(\lambda) - \int \log(\tau) dH(\tau),$$

which is nonzero in general except when $\mathbf{\Sigma}_p = \sigma^2 \mathbf{I}$. Moreover, similar arguments as in the case for the Frobenius loss show that

$$p^{-1} \sum_{i=1}^{p} \log(\mathbf{p}_{1i}^{\mathrm{T}} \mathbf{\Sigma}_p \mathbf{p}_{1i}) \mathbf{1}_{\{\lambda_{1i} \le x\}} \overset{a.s.}{\to} \int_{-\infty}^{x} \log(\delta_1(\lambda)) dF_1(\lambda)$$
$$= \int_{-\infty}^{x} \log(\delta(\lambda)) dF(\lambda).$$

Hence,

$$p^{-1} SL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\text{Ideal},m}) = p^{-1} \log\det(\widehat{\mathbf{\Sigma}}_{\text{Ideal},m} \mathbf{\Sigma}_p^{-1})$$
$$\overset{a.s.}{\to} \int \log(\delta_1(\lambda)) dF_1(\lambda) - \int \log(\tau) dH(\tau)$$
$$= \int \log(\delta(\lambda)) dF(\lambda) - \int \log(\tau) dH(\tau).$$

This completes the proof of the lemma. □

*Proof of Theorem 5.* We first assume (A1)', (A2)', (A3) and (A4). Since we have

$$\begin{aligned} p^{-1}\left\|\widehat{\boldsymbol{\Sigma}}_m - \boldsymbol{\Sigma}_p\right\|_F^2 &= p^{-1}\left\|\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{P}_1) - \mathbf{P}_1^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{P}_1\right\|_F^2 \\ &= p^{-1}\sum_{i=1}^{p}(\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i})^2 + p^{-1}\left\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{Ideal},m} - \boldsymbol{\Sigma}_p\right\|_F^2, \end{aligned}$$

we can express the efficiency loss in (4.2) with respect to the Frobenius loss as

$$\begin{aligned} EL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_m) = 1 - \Bigg( & \frac{p^{-1}\sum_{i=1}^{p}(\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i})^2}{p^{-1}\left\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{Ideal}} - \boldsymbol{\Sigma}_p\right\|_F^2} \\ & + \frac{p^{-1}\left\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{Ideal},m} - \boldsymbol{\Sigma}_p\right\|_F^2}{p^{-1}\left\|\widehat{\boldsymbol{\Sigma}}_{\mathrm{Ideal}} - \boldsymbol{\Sigma}_p\right\|_F^2} \Bigg)^{-1}. \end{aligned}$$

With $m/n \to 1$, we have $p/m \to c > 0$, same as $p/n$ does. The conclusion from lemma S.4 that the ratio of loss for the ideal estimators being almost surely 1 then implies that $EL(\widehat{\boldsymbol{\Sigma}}_m - \boldsymbol{\Sigma}_p) \overset{a.s.}{\to} 0$, if we can show that $p^{-1}\sum_{i=1}^{p}(\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i})^2 \overset{a.s.}{\to} 0$. But together with assumption $\sum_{n\geq 1} pn_2^{-5} < \infty$, we can apply Lemma 1, so that

$$\begin{aligned} p^{-1}\sum_{i=1}^{p}(\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i})^2 \leq & \left(\max_{1\leq i\leq p}\left|\frac{\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}}\right|\right)^2 \\ & \cdot \max_{1\leq i\leq p}(\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i})^2 \overset{a.s.}{\to} 0, \end{aligned}$$

since $\max_{1\leq i\leq p}(\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i})^2 \leq \left\|\boldsymbol{\Sigma}_p\right\|^2 = O(1)$ by assumption (A2)'. Hence, $EL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_m) \overset{a.s.}{\to} 0$ in this case.

Now consider the inverse Stein's loss. We have

$$\begin{aligned} SL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_m) &= \mathrm{tr}(\mathbf{P}_1^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{P}_1\mathrm{diag}^{-1}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{P}_1)) \\ &\quad - \log\det(\mathbf{P}_1^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{P}_1\mathrm{diag}^{-1}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{P}_1)) - p \\ &= \sum_{i=1}^{p}\frac{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i}} - p + \sum_{i=1}^{p}\log\left(\frac{\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}}\right) + SL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_{\mathrm{Ideal},m}) \\ &= \sum_{i=1}^{p}h_1(g_i) + \sum_{i=1}^{p}h_2(g_i) + SL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_{\mathrm{Ideal},m}), \end{aligned}$$

where we define $g_i$ for $i = 1, \ldots, p$, and the functions $h_1$ and $h_2$, which are differentiable on $(-1, \infty)$, to be

$$g_i = \frac{\mathbf{p}_{1i}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i}}, \quad h_1(x) = \frac{1}{1+x} - 1, \quad h_2(x) = \log(1+x).$$

With $SL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\text{Ideal},m})/SL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\text{Ideal}}) \overset{a.s.}{\to} 1$ by lemma S.4,

$$\begin{aligned} EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_m) = 1 - \Bigg( & \frac{p^{-1}\sum_{i=1}^p h_1(g_i)}{p^{-1}SL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\text{Ideal}})} \\ & + \frac{p^{-1}\sum_{i=1}^p h_2(g_1)}{p^{-1}SL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\text{Ideal}})} + \frac{p^{-1}SL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\text{Ideal},m})}{p^{-1}SL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{\text{Ideal}})} \Bigg)^{-1} \\ & \overset{a.s.}{\to} 0, \end{aligned}$$

if we can show further that $p^{-1}\sum_{i=1}^p h_j(g_i) \overset{a.s.}{\to} 0$ for $j = 1, 2$. To this end, by Lemma 1, $g_i$ is almost surely 0 for all $i$. Hence, in particular, for $n$ large enough, we have $|g_i| \leq B < 1$ almost surely, so that $h_j(g_i)$ is well defined for all $i$. This implies that for $j = 1, 2$,

$$p^{-1}\sum_{i=1}^p h_j(g_i) \leq \max_{1\leq i\leq p} h_j(g_i) = \max_{1\leq i\leq p} |g_i||h_j'(\eta_i)| \overset{a.s.}{\to} 0,$$

where $\eta_i$ lies almost surely in $[-B, B]$ for all $i$ for large enough $n$, so that the $h_j'(\eta_i)$'s are bounded almost surely for all $i$. This completes the proof of the theorem. □

*Proof of Theorem 6.* For the Frobenius loss,

$$\begin{aligned} \big\|\widehat{\mathbf{\Sigma}}_{m,M} - \mathbf{\Sigma}_p\big\|_F^2 = \Big\|\frac{1}{M}\sum_{i=1}^M (\widehat{\mathbf{\Sigma}}_m^{(i)} - \mathbf{\Sigma}_p)\Big\|_F^2 & \leq \Big(\frac{1}{M}\sum_{i=1}^M \big\|\widehat{\mathbf{\Sigma}}_m^{(i)} - \mathbf{\Sigma}_p\big\|_F\Big)^2 \\ & \leq \frac{1}{M}\sum_{i=1}^M \big\|\widehat{\mathbf{\Sigma}}_m^{(i)} - \mathbf{\Sigma}_p\big\|_F^2, \end{aligned}$$

so that

$$\begin{aligned} EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_{m,M}) \leq 1 - \frac{\big\|\widehat{\mathbf{\Sigma}}_{\text{Ideal}} - \mathbf{\Sigma}_p\big\|_F^2}{\frac{1}{M}\sum_{i=1}^M \big\|\widehat{\mathbf{\Sigma}}_m^{(i)} - \mathbf{\Sigma}_p\big\|_F^2} = 1 - \frac{1}{\frac{1}{M}\sum_{i=1}^M \frac{1}{1-EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_m^{(i)})}} \\ \overset{a.s.}{\to} 0, \end{aligned}$$

since $EL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}_m^{(i)}) \overset{a.s.}{\to} 0$ by Theorem 5. This completes the proof for the Frobenius loss.

For the inverse Stein's loss, note that for any estimator $\widehat{\mathbf{\Sigma}}$,

$$SL(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}) = \operatorname{tr}\Big((\mathbf{\Sigma}_p - \widehat{\mathbf{\Sigma}})\widehat{\mathbf{\Sigma}}^{-1}\Big) + \log\det(\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}_p^{-1}).$$

Hence

$$\begin{aligned}
SL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_{m,M}) &= \frac{1}{M}\sum_{i=1}^{M} \mathrm{tr}\Big((\boldsymbol{\Sigma}_p - \widehat{\boldsymbol{\Sigma}}_m^{(i)})\widehat{\boldsymbol{\Sigma}}_{m,M}^{-1}\Big) + \log\det\Big(\frac{1}{M}\sum_{i=1}^{M}\widehat{\boldsymbol{\Sigma}}_m^{(i)}\boldsymbol{\Sigma}_p^{-1}\Big) \\
&\leq \frac{1}{M}\sum_{i=1}^{M} \mathrm{tr}\Big((\boldsymbol{\Sigma}_p - \widehat{\boldsymbol{\Sigma}}_m^{(i)})\widehat{\boldsymbol{\Sigma}}_{m,M}^{-1}\Big) + \frac{1}{M}\sum_{i=1}^{M}\log\det(\widehat{\boldsymbol{\Sigma}}_m^{(i)}\boldsymbol{\Sigma}_p^{-1}) \\
&= \frac{1}{M}\sum_{i=1}^{M} \mathrm{tr}\Big((\boldsymbol{\Sigma}_p - \widehat{\boldsymbol{\Sigma}}_m^{(i)})\big(\widehat{\boldsymbol{\Sigma}}_{m,M}^{-1} - (\widehat{\boldsymbol{\Sigma}}_m^{(i)})^{-1}\big)\Big) + \frac{1}{M}\sum_{i=1}^{M} SL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_m^{(i)}),
\end{aligned}$$

where the second line comes from the convexity of $\log\det(\mathbf{X})$ where $\mathbf{X}$ is positive semi-definite. Then

$$\begin{aligned}
EL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_{m,M}) \leq 1 - \Bigg( & \frac{\frac{1}{pM}\sum_{i=1}^{M} \mathrm{tr}\Big((\boldsymbol{\Sigma}_p - \widehat{\boldsymbol{\Sigma}}_m^{(i)})\big(\widehat{\boldsymbol{\Sigma}}_{m,M}^{-1} - (\widehat{\boldsymbol{\Sigma}}_m^{(i)})^{-1}\big)\Big)}{p^{-1}SL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_{\mathrm{Ideal}})} \\
& + \frac{1}{M}\sum_{i=1}^{M}\frac{1}{1 - EL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_m^{(i)})}\Bigg)^{-1} \\
& \overset{a.s.}{\to} 0,
\end{aligned}$$

if we can show further that $\frac{1}{pM}\sum_{i=1}^{M} \mathrm{tr}\Big((\boldsymbol{\Sigma}_p - \widehat{\boldsymbol{\Sigma}}_m^{(i)})\big(\widehat{\boldsymbol{\Sigma}}_{m,M}^{-1} - (\widehat{\boldsymbol{\Sigma}}_m^{(i)})^{-1}\big)\Big) \overset{a.s.}{\to} 0$, since by Theorem 5 we have $EL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_m^{(i)}) \overset{a.s.}{\to} 0$ for each $i$, and that in the proof of Lemma S.4 it is proved that $p^{-1}SL(\boldsymbol{\Sigma}_p, \widehat{\boldsymbol{\Sigma}}_{\mathrm{Ideal}})$ converges almost surely to a nonzero value when $\boldsymbol{\Sigma}_p \neq \sigma^2\mathbf{I}$. For this, write the term as $D_1 + D_2$, where

$$D_1 = \frac{1}{pM}\sum_{i=1}^{M}\mathrm{tr}\big((\widehat{\boldsymbol{\Sigma}}_m^{(i)} - \boldsymbol{\Sigma}_p)(\widehat{\boldsymbol{\Sigma}}_m^{(i)})^{-1}\big),\ D_2 = \frac{1}{pM}\sum_{i=1}^{M}\mathrm{tr}\big((\boldsymbol{\Sigma}_p - \widehat{\boldsymbol{\Sigma}}_m^{(i)})\widehat{\boldsymbol{\Sigma}}_{m,M}^{-1}\big).$$

For $D_1$, writing $\mathbf{P}_{1i} = (\mathbf{p}_{1i,1}, \ldots, \mathbf{p}_{1i,p})$, we can use Lemma 1 to deduce that

$$\begin{aligned}
D_1 &= \frac{1}{pM}\sum_{i=1}^{M}\mathrm{tr}\big((\mathrm{diag}(\mathbf{P}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2^{(i)}\mathbf{P}_{1i}) - \mathbf{P}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{P}_{1i})\mathrm{diag}^{-1}(\mathbf{P}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2^{(i)}\mathbf{P}_{1i})\big) \\
&= \frac{1}{pM}\sum_{i=1}^{M}\sum_{j=1}^{p}\frac{\mathbf{p}_{1i,j}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2^{(i)}\mathbf{p}_{1i,j} - \mathbf{p}_{1i,j}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i,j}}{\mathbf{p}_{1i,j}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2^{(i)}\mathbf{p}_{1i,j}} \\
&\leq \frac{1}{M}\sum_{i=1}^{M}\max_{1\leq j\leq p}\left|\frac{\mathbf{p}_{1i,j}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2^{(i)}\mathbf{p}_{1i,j} - \mathbf{p}_{1i,j}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i,j}}{\mathbf{p}_{1i,j}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i,j}}\right| \\
&\quad \cdot\left(1 - \max_{1\leq j\leq p}\left|\frac{\mathbf{p}_{1i,j}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2^{(i)}\mathbf{p}_{1i,j} - \mathbf{p}_{1i,j}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i,j}}{\mathbf{p}_{1i,j}^{\mathrm{T}}\boldsymbol{\Sigma}_p\mathbf{p}_{1i,j}}\right|\right)^{-1} \overset{a.s.}{\to} 0.
\end{aligned}$$

For $D_2$, define $\widehat{\mathbf{\Sigma}}_{\mathrm{CRC}}^{(i)} = \mathbf{P}_{1i}\Big(\frac{1}{M}\sum_{j=1}^{M}\mathrm{diag}(\mathbf{P}_{1j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{P}_{1j})\Big)\mathbf{P}_{1i}$, and define also $\mathbf{B}_i = (\widehat{\mathbf{\Sigma}}_{\mathrm{CRC}}^{(i)} - \widehat{\mathbf{\Sigma}}_{m,M})(\widehat{\mathbf{\Sigma}}_{\mathrm{CRC}}^{(i)})^{-1}$. Then

$$\widehat{\mathbf{\Sigma}}_{m,M}^{-1} = (\widehat{\mathbf{\Sigma}}_{\mathrm{CRC}}^{(i)})^{-1}(\mathbf{I}-\mathbf{B}_i)^{-1} = (\widehat{\mathbf{\Sigma}}_{\mathrm{CRC}}^{(i)})^{-1} + (\widehat{\mathbf{\Sigma}}_{\mathrm{CRC}}^{(i)})^{-1}\sum_{k\geq 1}\mathbf{B}_i^k. \tag{S.4}$$

Suppose $\max_{1\leq i,j\leq M}\big\|\mathbf{P}_{1i}-\mathbf{P}_{1j}\big\| \overset{a.s.}{\to} 0$, to be proved at the end of this proof. Then the above Neumann's series expansion is valid, since then for large enough $n, p$ with $p/n \to c > 0$,

$$\begin{aligned}
\big\|\mathbf{B}_i\big\| &\leq \frac{1}{M}\Bigg(2\Big\|\sum_{j=1}^{M}(\mathbf{P}_{1i}-\mathbf{P}_{1j})\mathrm{diag}(\mathbf{P}_{1j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{P}_{1j})\mathbf{P}_{1i}^{\mathrm{T}}\Big\| \\
&\qquad + \Big\|\sum_{j=1}^{M}(\mathbf{P}_{1i}-\mathbf{P}_{1j})\mathrm{diag}(\mathbf{P}_{1j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{P}_{1j})(\mathbf{P}_{1i}-\mathbf{P}_{1j})^{\mathrm{T}}\Big\|\Bigg) \\
&\qquad \cdot \Big(\min_{1\leq j\leq M}\min_{1\leq k\leq p}\mathbf{p}_{1j,k}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{p}_{1j,k}\Big)^{-1} \\
&\leq \max_{1\leq j\leq M}\big\|\mathbf{P}_{1i}-\mathbf{P}_{1j}\big\|\max_{1\leq k\leq p}\mathbf{p}_{1j,k}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1j,k}\Big(2+\big\|\mathbf{P}_{1i}-\mathbf{P}_{1j}\big\|\Big) \\
&\qquad \cdot \Big(\min_{1\leq j\leq M}\min_{1\leq k\leq p}\mathbf{p}_{1j,k}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1j,k}\Big)^{-1} \\
&\leq \max_{1\leq i,j\leq M}\big\|\mathbf{P}_{1i}-\mathbf{P}_{1j}\big\|\big\|\mathbf{\Sigma}_p\big\|(2+\big\|\mathbf{P}_{1i}-\mathbf{P}_{1j}\big\|)/\lambda_{\min}(\mathbf{\Sigma}_p) \overset{a.s.}{\to} 0,
\end{aligned} \tag{S.5}$$

where the second last line follows from Lemma 1 when $n, p$ is large. With (S.4), $D_2 = D_3 + D_4$, where

$$\begin{aligned}
D_3 &= \frac{1}{pM}\sum_{i=1}^{M}\mathrm{tr}\Big((\mathbf{\Sigma}_p - \widehat{\mathbf{\Sigma}}_m^{(i)})(\widehat{\mathbf{\Sigma}}_{\mathrm{CRC}}^{(i)})^{-1}\Big), \\
D_4 &= \frac{1}{pM}\sum_{i=1}^{M}\mathrm{tr}\Bigg((\mathbf{\Sigma}_p - \widehat{\mathbf{\Sigma}}_m^{(i)})(\widehat{\mathbf{\Sigma}}_{\mathrm{CRC}}^{(i)})^{-1}\sum_{k\geq 1}\mathbf{B}_i^k\Bigg).
\end{aligned}$$

Similar to the analysis of $D_1$, using Lemma 1, we have

$$\begin{aligned}
D_3 &= \frac{1}{pM}\sum_{i=1}^{M}\sum_{k=1}^{p}\frac{\mathbf{p}_{1i,k}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i,k} - \mathbf{p}_{1i,k}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(i)}\mathbf{p}_{1i,k}}{\frac{1}{M}\sum_{j=1}^{M}\mathbf{p}_{1j,k}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{p}_{1j,k}} \\
&\leq \max_{1\leq i\leq M}\max_{1\leq k\leq p}\left|\frac{\mathbf{p}_{1i,k}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i,k} - \mathbf{p}_{1i,k}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(i)}\mathbf{p}_{1i,k}}{\mathbf{p}_{1i,k}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i,k}}\right|\cdot\mathbf{p}_{1i,k}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1i,k} \\
&\qquad \cdot\Bigg(\min_{1\leq j\leq M}\min_{1\leq k\leq p}\left(\frac{\mathbf{p}_{1j,k}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{p}_{1j,k} - \mathbf{p}_{1j,k}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1j,k}}{\mathbf{p}_{1j,k}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1j,k}}\right)\mathbf{p}_{1j,k}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1j,k} \\
&\qquad + \mathbf{p}_{1j,k}^{\mathrm{T}}\mathbf{\Sigma}_p\mathbf{p}_{1j,k}\Bigg)^{-1} \overset{a.s.}{\to} 0.
\end{aligned}$$

For $D_4$, with $\|\mathbf{B}_i\| \overset{a.s.}{\to} 0$ proved in (S.5) and using Lemma 1, we have

$$\begin{aligned} D_4 &\leq \max_{1\leq i\leq M} \|\widehat{\boldsymbol{\Sigma}}_m^{(i)} - \boldsymbol{\Sigma}_p\| \cdot \lambda_{\min}^{-1}(\widehat{\boldsymbol{\Sigma}}_{\mathrm{CRC}}^{(i)}) \cdot \sum_{k\geq 1} \|\mathbf{B}_i\|^k \\ &\leq \frac{(\|\boldsymbol{\Sigma}_p\| + \max_{1\leq i\leq M} \max_{1\leq k\leq p} \mathbf{p}_{1i,k}^{\mathrm{T}} \widetilde{\boldsymbol{\Sigma}}_2^{(i)} \mathbf{p}_{1i,k}) \|\mathbf{B}_i\|}{\min_{1\leq i\leq M} \min_{1\leq k\leq p} \mathbf{p}_{1i,k}^{\mathrm{T}} \widetilde{\boldsymbol{\Sigma}}_2^{(i)} \mathbf{p}_{1i,k} (1 - \|\mathbf{B}_i\|)} \overset{a.s.}{\to} 0. \end{aligned}$$

Hence if we can show further that $\max_{1\leq i,j\leq M} \|\mathbf{P}_{1i} - \mathbf{P}_{1j}\| \overset{a.s.}{\to} 0$, the proof of the Theorem is complete. To this end, using (S.3), we have for any $1 \leq i \leq M$, with $\lambda_{1i,k}$ denoting the $k$th largest eigenvalue of $\widetilde{\boldsymbol{\Sigma}}_1^{(i)}$,

$$p^{-1} \sum_{k=1}^{p} g(\mathbf{p}_{1i,k}^{\mathrm{T}} \boldsymbol{\Sigma}_p \mathbf{p}_{1i,k}) \mathbf{1}_{\{\lambda_{1i,k} \leq x\}} \overset{a.s.}{\to} \int_{-\infty}^{x} g(\delta(\lambda)) dF(\lambda),$$

since $m/n \to 1$ implies that $p/n, p/m \to c > 0$. Setting $g \equiv 1$, it is easy to see that $\lambda_{1i,k}$ is almost surely the same as $\lambda_{1j,k}$ for $1 \leq i, j \leq M$ and $1 \leq k \leq p$, as $n, p \to \infty$. These imply that for $1 \leq i \leq M$, $\mathbf{p}_{1i,k}^{\mathrm{T}} \boldsymbol{\Sigma}_p \mathbf{p}_{1i,k}$ is almost surely the same as $\delta(\lambda_{11,k})$, and hence for $1 \leq i, j \leq M$ and $1 \leq k \leq p$,

$$\mathbf{p}_{1i,k}^{\mathrm{T}} \boldsymbol{\Sigma}_p \mathbf{p}_{1i,k} - \mathbf{p}_{1j,k}^{\mathrm{T}} \boldsymbol{\Sigma}_p \mathbf{p}_{1j,k} \overset{a.s.}{\to} 0.$$

But $\mathbf{p}_{1i,k}^{\mathrm{T}} \boldsymbol{\Sigma}_p \mathbf{p}_{1i,k} - \mathbf{p}_{1j,k}^{\mathrm{T}} \boldsymbol{\Sigma}_p \mathbf{p}_{1j,k} = (\mathbf{p}_{1i,k} - \mathbf{p}_{1j,k})^{\mathrm{T}} \boldsymbol{\Sigma}_p (\mathbf{p}_{1i,k} + \mathbf{p}_{1j,k})$, hence if $\mathbf{p}_{1i,k}$ or $\mathbf{p}_{1j,k}$ are not eigenvectors of $\boldsymbol{\Sigma}_p$ asymptotically (this excludes the case $\boldsymbol{\Sigma}_p = \sigma^2 \mathbf{I}$, which by assumption is not the case), then there exists a constant $a > 0$ such that

$$(\mathbf{p}_{1i,k} - \mathbf{p}_{1j,k})^{\mathrm{T}} \boldsymbol{\Sigma}_p (\mathbf{p}_{1i,k} + \mathbf{p}_{1j,k}) = \|\mathbf{p}_{1i,k} - \mathbf{p}_{1j,k}\| \cdot a \|\mathbf{p}_{1i,k} + \mathbf{p}_{1j,k}\| \overset{a.s.}{\to} 0,$$

implying that for all $1 \leq i, j \leq M$, $1 \leq k \leq p$, $\|\mathbf{p}_{1i,k} \pm \mathbf{p}_{1j,k}\| \overset{a.s.}{\to} 0$. With the correct orientation chosen, we must then have $\max_{1\leq i,j\leq M} \|\mathbf{P}_i - \mathbf{P}_j\| \overset{a.s.}{\to} 0$.

Finally, if both $\mathbf{p}_{1i,k}$ and $\mathbf{p}_{1j,k}$ are asymptotic eigenvectors of $\boldsymbol{\Sigma}_p$ for some $1 \leq k \leq p$, then assuming $\|\boldsymbol{\Sigma}_p \mathbf{p}_{1i,k} - \lambda \mathbf{p}_{1i,k}\| \overset{a.s.}{\to} 0$ and $\|\boldsymbol{\Sigma}_p \mathbf{p}_{1j,k} - \gamma \mathbf{p}_{1j,k}\| \overset{a.s.}{\to} 0$,

$$\begin{aligned} (\mathbf{p}_{1i,k} - \mathbf{p}_{1j,k})^{\mathrm{T}} \boldsymbol{\Sigma}_p (\mathbf{p}_{1i,k} + \mathbf{p}_{1j,k}) &= (\mathbf{p}_{1i,k} - \mathbf{p}_{1j,k})^{\mathrm{T}} (\lambda \mathbf{p}_{1i,k} + \gamma \mathbf{p}_{1j,k}) \\ &\quad + (\mathbf{p}_{1i,k} - \mathbf{p}_{1j,k})^{\mathrm{T}} (\boldsymbol{\Sigma}_p \mathbf{p}_{1i,k} - \lambda \mathbf{p}_{1i,k} \\ &\quad + \boldsymbol{\Sigma}_p \mathbf{p}_{1j,k} - \gamma \mathbf{p}_{1j,k}) \\ &= (\lambda - \gamma)(1 - \mathbf{p}_{1i,k}^{\mathrm{T}} \mathbf{p}_{1j,k}) + o(1), \end{aligned}$$

so that if $\lambda \neq \gamma$, we must have $\|\mathbf{p}_{1i,k} - \mathbf{p}_{1j,k}\| \overset{a.s.}{\to} 0$, which is what we want. If $\lambda = \gamma$, then $\mathbf{p}_{1i,k}$ and $\mathbf{p}_{1j,k}$ are asymptotically spanning the same eigenspace, hence $\max_{1\leq i,j\leq M} \|\mathbf{P}_i - \mathbf{P}_j\| \overset{a.s.}{\to} 0$ still holds. This completes the proof of the theorem. □

# References


Abadir, K. M., Distaso, W., and Žikeš, F. (2014). Design-free estimation of variance matrices. *Journal of Econometrics*, 181(2):165 – 180.

Bai, Z. and Silverstein, J. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics, New York, 2 edition.

Bai, Z. D. and Silverstein, J. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1):316–345.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060.

Ledoit, O. and Wolf, M. (2013a). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. ECON - Working Papers 122, Department of Economics - University of Zurich.

Ledoit, O. and Wolf, M. (2013b). Spectrum estimation: a unified framework for covariance matrix estimation and PCA in large dimensions. ECON - Working Papers 105, Department of Economics - University of Zurich.

Silverstein, J. and Bai, Z. (1995). On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175 – 192.