Optimal Uncertainty Size in Distributionally Robust Inverse Covariance Estimation

Jose Blanchet  Nian Si

Management Science and Engineering, Stanford University

Abstract

In a recent paper, Nguyen, Kuhn, and Esfahani (2018) built a distributionally robust estimator for the precision matrix (i.e. inverse covariance matrix) of a multivariate Gaussian distribution. The distributional uncertainty size is a key ingredient in the construction of such estimator, which is shown to have an excellent empirical performance. In this paper, we develop a statistical theory which shows how to optimally choose the uncertainty size to minimize the associated Stein loss. Surprisingly, the optimal uncertainty size scales linearly with the sample size, instead of the canonical square-root scaling which may be expected for this problem.

1 Introduction

Motivated by a wide range of problems which require the estimation of the inverse of a covariance matrix recently, [7] constructed an estimator based on distributionally robust optimization using the Wasserstein distance in Euclidean space. A crucial ingredient is the distributional uncertainty size, which plays the role of a regularization parameter.

In their paper, [7] show excellent empirical performance of their estimator in comparison to several estimators (based on shrinkage and regularization) used in practice. The comparison is based in terms of the corresponding Stein loss (defined in terms of the likelihood, as we shall review). However, no theory is provided in order to choose the distributional uncertainty size.

Our goal is to provide an asymptotically optimal expression for the distributional uncertainty size, in terms of the Stein loss performance, as the sample size increases.

Our development provides interesting insights which validate the empirical observations in [7]. In particular, in the Introduction of [7], leading to their equation (4), they argue that the distributional uncertainty size, $\rho_n$, should scale at rate $\rho_n = O\left(n^{-1/2}\right)$ (where $n$ is the sample size) due to the existence of a Central Limit Theorem for the Wasserstein distance for Gaussian distributions. However, the numerical experiments, reported in Section 6.1 of [7], suggest an optimal scaling of the form $\rho_n = O\left(n^{-\kappa}\right)$ where $\kappa > 1/2$.

Our main result shows that the asymptotically optimal choice of distributional uncertainty is of the form $\rho_n = \rho_\ast n^{-1}(1 + o(1))$ as $n \to \infty$ where $\rho_\ast > 0$ is a constant which is characterized explicitly. Our results therefore validate the empirical findings of [7] with $\kappa = 1.$
We will review the estimator of \[7\] and state our main result in Section 2. Then, we will provide the proof of our result in Section 3. Numerical experiments are included which provide a sense of the non-asymptotic performance of our asymptotically optimal choice, the results (non-surprisingly) validate our theoretical findings; these experiments are reported in Section 4.

2 Basic Notions and Main Result

We now review the basic definitions underlying the estimator from \[7\]. Suppose we have i.i.d. samples \(\xi_i \sim \mathcal{N}(0, \Sigma_0)\) (normally distributed with mean zero and covariance matrix \(\Sigma_0\)), where \(\xi_i \in \mathbb{R}^d\) and \(\Sigma_0\) is assumed to be strictly positive definite. We write

\[
\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^{n} \xi_i \xi_i^T,
\]
and let \(\hat{\mathbb{P}}_n\) correspond to a distribution with mean zero and covariance matrix \(\hat{\Sigma}_n\), which we denote as \(\mathcal{N}(0, \hat{\Sigma}_n)\). Throughout our development we use the notation \(\langle A, B \rangle = \text{tr}(A^T B)\) for any \(d \times d\) matrices \(A, B\), where \(A^T\) denotes the transpose of \(A\).

We define the Stein loss as

\[
L(X, \Sigma_0) = -\log \det(X \Sigma_0) + \langle X, \Sigma_0 \rangle - d,
\]
where \(X\) is any estimator of the precision matrix (i.e. the inverse covariance matrix).

Given an uncertainty size \(\rho\), let us write \(X_n^*(\rho)\) for the distributionally robust estimator proposed in \[7\]; i.e.,

\[
X_n^*(\rho) = \arg \min_{X > 0} \left\{ -\log \det X + \sup_{Q \in \mathcal{P}_\rho} \mathbb{E}^Q \left[ \langle \xi \xi^T, X \rangle \right] \right\},
\]
where \(\mathcal{P}_\rho\) is the set of \(d\)-dimensional normal distributions with mean zero and which lie within distance \(\rho\) measured in Wasserstein sense, which we define next; see, for example, Chapter 7 in [9] for background on Wasserstein distances and, more generally, optimal transport costs. The Wasserstein distance (more precisely, the Wasserstein distance of order two with Euclidean norm) is defined as follows. First, let \(\mathcal{M}_+(\mathbb{R}^d \times \mathbb{R}^d)\) be the set of Borel (positive) measures on \(\mathbb{R}^d \times \mathbb{R}^d\) and define the Wasserstein distance between \(\hat{\mathbb{P}}_n\) and \(\mathbb{Q}\) via

\[
W_2(\hat{\mathbb{P}}_n, \mathbb{Q}) = \inf_{\pi \in \mathcal{M}_+(\mathbb{R}^d \times \mathbb{R}^d)} \left\{ \left( \int \|z - w\|^2 \pi (dz, dw) \right)^{1/2} \right\},
\]

\[
: \int_{w \in \mathbb{R}^d} \pi (dx, dw) = \hat{\mathbb{P}}_n (dx), \int_{x \in \mathbb{R}^d} \pi (dx, dw) = \mathbb{Q}(dw) \}\.
\]

Then,

\[
\mathcal{P}_\rho = \left\{ \mathbb{Q} \sim \mathcal{N}(0, \Sigma) \text{ for some } \Sigma : W_2(\hat{\mathbb{P}}_n, \mathbb{Q}) \leq \rho \right\}.
\]
In simple terms, \( \mathcal{P}_\rho \) is the set of probability measures corresponding to a Gaussian distribution which lie within \( \rho \) units in Wasserstein distance from \( \hat{\mathcal{P}}_n \). It is well known (in fact, an immediate consequence of the delta method) that \( n^{1/2} W_2(\hat{\mathcal{P}}_n, \mathcal{P}_\infty) \Rightarrow \mathcal{W} \) for some limit law \( \mathcal{W} \) which can be explicitly characterized (but not important for our development; see [8]). This result suggests that \( \rho := \rho_n \) should scale in order \( O \left( n^{-1/2} \right) \). It is therefore somewhat surprising that the optimal scaling of \( \rho \) for the purpose of minimizing the Stein loss is actually significantly smaller, as the main result of this paper indicates next.

**Theorem 1.** Let

\[
\rho_n = \arg\min_{\rho \geq 0} \{ E[L(X^*_n(\rho), \Sigma_0)] \},
\]

then

\[
n\rho_n \to \rho_*,
\]

for \( \rho_* > 0 \).

**Remark:** The explicit expression of \( \rho_* \) can be characterized as follows. First, let consider the weak limit

\[
Z = \lim_{n \to \infty} n^{1/2} \left( \hat{\Sigma}_n - \Sigma_0 \right),
\]

which, by the Central Limit Theorem is a matrix with correlated mean zero Gaussian entries. Then, we have that

\[
\rho_* = E \left( \frac{4 \text{tr} (\Sigma^{-2}_0 Z \Sigma^{-1}_0 Z)}{\text{tr}(\Sigma^{-1}_0)^{1/2}} - \frac{\text{tr}(Z \Sigma^{-2}_0 Z)^2}{\text{tr}(\Sigma^{-1}_0)^{3/2}} \right) \frac{\text{tr}(\Sigma^{-1}_0)}{4 \text{tr}(\Sigma^{-2}_0)}.
\]

Theorem [1] indicates that \( \rho_* > 0 \), which will be verified as a part of the proof of this result.

### 3 Proof of Theorem [1]

We first collect the following observations, which we summarize in the form of propositions and lemmas for which we provide references or corresponding proofs in the appendix. Then, we use these results to develop the proof of Theorem [1].

#### 3.1 Auxiliary Results

First, we provide a lemma based on the analytical solution (Theorem 3.1 in [2]).

**Lemma 1.** When \( n > d \), with probability 1, we have

\[
X^*_n(\rho) = \hat{\Sigma}^{-1}_n + \hat{A}_n \rho + O(\rho^2),
\]

\[
X^*_n(\rho)^{-1} = \hat{\Sigma}_n - \hat{\Sigma}_n \hat{A}_n \hat{\Sigma}_n \rho + O(\rho^2).
\]

where

\[
\hat{A}_n = -\frac{2}{\sqrt{\text{tr}(\hat{\Sigma}^{-1}_n)}} \hat{\Sigma}^{-2}_n.
\]
and the constant in the errors involving $O(\rho^2)$ are polynomial in the ratio of the maximum to the minimum eigenvalues of $\hat{\Sigma}_n$.

From Lemma 1 we have that

$$X_n^*(\rho_{\text{n}})^{-1} - \Sigma_0 = \left(\hat{\Sigma}_n - \Sigma_0\right) - \hat{\Sigma}_n \hat{A}_n \hat{\Sigma}_n \rho + O(\rho^2).$$

(2)

The first proposition provides standard asymptotic normality results for various estimators (see, for example, Chapter 1 in [6]).

**Proposition 1.** The following convergence results hold

1. $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \Rightarrow N(0, \Sigma_0)$,
2. $\sqrt{n} \left(\hat{\Sigma}_n - \Sigma_0\right) \Rightarrow Z$, where $Z$ is a matrix of jointly Gaussian random variables with mean zero and

$$\text{cov}(Z_{i_1,j_1}, Z_{i_2,j_2}) = \mathbb{E} \xi^{(i_1)} \xi^{(j_1)} \xi^{(i_2)} \xi^{(j_2)} - \left(\mathbb{E} \xi^{(i_1)} \xi^{(j_1)}\right) \left(\mathbb{E} \xi^{(i_2)} \xi^{(j_2)}\right)$$

(3)

$$= \sigma^2_{i_1,i_2} \sigma^2_{j_1,j_2} + \sigma^2_{i_1,j_2} \sigma^2_{j_1,i_2},$$

(4)

where $\xi^{(i)}$ is the $i$-th entry of $\xi$ and $\sigma^2_{i,j} = \text{cov}(\xi^{(i)}, \xi^{(j)})$.

3. $\hat{A}_n \Rightarrow A_0$ and $\sqrt{n} \left(\hat{A}_n - A_0\right) \Rightarrow Z_A$, where $A_0 = -\frac{2}{\sqrt{\text{tr}(\Sigma_0^{-1})}} \Sigma_0^{-2}$ and

$$Z_A = -\frac{\text{tr}(\Sigma_0^{-1} Z \Sigma_0^{-1}) \Sigma_0^{-2}}{\text{tr}(\Sigma_0^{-1})^{3/2}} + 2 \frac{\Sigma^{-1} Z \Sigma^{-2} + \Sigma^{-2} Z \Sigma^{-1}}{\text{tr}(\Sigma^{-1})^{1/2}}.$$

Further, we also have the following observations.

**Proposition 2.** We have that $\mathbb{E} \langle Z, Z_A \rangle > 0$.

**Lemma 2.**

i) $\mathbb{E} \left[\langle \hat{\Sigma}_n \hat{A}_n \hat{\Sigma}_n, \hat{A}_n \rangle\right] \rightarrow \langle \Sigma_0 A_0 \Sigma_0, A_0 \rangle$.

ii) $\mathbb{E} \left<\sqrt{n} \left(\hat{\Sigma}_n - \Sigma_0\right), \hat{A}_n\right> \rightarrow \mathbb{E} \langle Z, A_0 \rangle$.

iii) $\mathbb{E} \left<\sqrt{n} \left(\hat{\Sigma}_n - \Sigma_0\right), \sqrt{n} \left(\hat{A}_n - A_0\right)\right> \rightarrow \mathbb{E} \langle Z, Z_A \rangle$.

Using the previous technical results we are ready to provide the proof of Theorem 1.

### 3.2 Development of Proof of Theorem 1

The gradient of $L(\cdot, \Sigma_0)$ satisfies

$$h(X, \Sigma_0) = \frac{\partial L(X, \Sigma_0)}{\partial X} = -X^{-1} + \Sigma_0.$$

and

$$\frac{\partial L(X^*_n(\rho_{\text{n}}), \Sigma_0)}{\partial \rho} = \left<h(X^*_n(\rho_{\text{n}}), \Sigma_0), \hat{A}_n\right> = 0.$$
We want to choose \( \rho_n \) that minimizes \( \mathbb{E} [ L(X^*_n(\rho_n), \Sigma_0) ] \). Due to the convexity of loss function, the problem is equivalent to
\[
\mathbb{E} \left< h(X^*_n(\rho_n), \Sigma_0), \hat{A}_n \right> = 0. \tag{5}
\]

By plugging (2) into (5), we have
\[
\mathbb{E} \left< h(X^*_n(\rho_n), \Sigma_0), \hat{A}_n \right> = -\mathbb{E} \left< \hat{\Sigma}_n - \Sigma_0 - \hat{\Sigma}_n \hat{A}_n \hat{\Sigma}_n \rho_n + O(\rho_n^2), \hat{A}_n \right> = 0, \tag{6}
\]
which is equivalent to
\[
\mathbb{E} \left[ \left< \hat{\Sigma}_n \hat{A}_n \hat{\Sigma}_n, \hat{A}_n \right> \right] \rho_n + O(\rho_n^2) = \mathbb{E} \left< \hat{\Sigma}_n - \Sigma_0, \hat{A}_n \right>. \tag{7}
\]
The validity of expanding the expectations follows by applying the uniform integrability results underlying the proof of Lemma 2.

Now, note also by Lemma 2,
\[
\lim_{n \to \infty} \mathbb{E} \left[ \left< \hat{\Sigma}_n \hat{A}_n \hat{\Sigma}_n, \hat{A}_n \right> \right] = \langle \Sigma_0 A_0 \Sigma_0, A_0 \rangle = 4 \text{tr}(\Sigma_0^{-2})/\text{tr}(\Sigma_0^{-1}) > 0.
\]

By multiplying \( \sqrt{n} \) on both sides of (7) and by Slutsky’s lemma (Theorem 1.8.10 in [6]), we have
\[
\lim_{n \to 0} \sqrt{n} \left( \mathbb{E} \left[ \left< \hat{\Sigma}_n \hat{A}_n \hat{\Sigma}_n, \hat{A}_n \right> \right] \rho_n + O(\rho_n^2) \right) = \lim_{n \to \infty} \mathbb{E} \left< \sqrt{n} \left( \hat{\Sigma}_n - \Sigma_0 \right), \hat{A}_n \right> = \mathbb{E} \langle Z, A_0 \rangle = 0.
\]

Therefore,
\[
\lim_{n \to \infty} \sqrt{n} \rho_n = 0.
\]

Furthermore, since \( \mathbb{E} \left[ \hat{\Sigma}_n - \Sigma_0 \right] = 0 \) for every \( n \), we have (once again by Lemma 2)
\[
\lim_{n \to \infty} \mathbb{E} \left< n \left( \hat{\Sigma}_n - \Sigma_0 \right), \hat{A}_n \right> = \lim_{n \to \infty} \mathbb{E} \left< n \left( \hat{\Sigma}_n - \Sigma_0 \right), \hat{A}_n - A_0 + A_0 \right> = \lim_{n \to \infty} \left< n \mathbb{E} \left[ \hat{\Sigma}_n - \Sigma_0 \right], A_0 \right> + \lim_{n \to \infty} \mathbb{E} \left< \sqrt{n} \left( \hat{\Sigma}_n - \Sigma_0 \right), \sqrt{n} \left( \hat{A}_n - A_0 \right) \right> = \mathbb{E} \langle Z, Z A \rangle .
\]

By multiplying \( n \) on sides of (7), we have
\[
\lim_{n \to \infty} n \rho_n = \rho^* = \frac{\mathbb{E} \langle Z, Z A \rangle}{\langle \Sigma_0 A_0 \Sigma_0, A_0 \rangle} = \mathbb{E} \left( \frac{4 \text{tr}(\Sigma_0^{-2} Z \Sigma_0^{-1} Z)}{\text{tr}(\Sigma_0^{-1})^{1/2}} - \frac{\text{tr}(Z \Sigma_0^{-2})^2}{\text{tr}(\Sigma_0^{-1})^{3/2}} \right) \frac{\text{tr}(\Sigma_0^{-1})}{4 \text{tr}(\Sigma_0^{-2})} > 0,
\]
and the result follows.

4 Numerical Experiments

We provide various numerical experiments to provide an empirical validation of our theory and the performance of the asymptotically optimal choice of uncertainty size in finite samples.
The first example is in one dimension. The data is sampled from a normal distribution, \( N(0, \sigma_0^2) \); i.e., \( \Sigma_0 = \sigma_0^2 \) in the real line. Therefore,

\[
A_0 = -2\sigma_0^{-3}, \quad \mathbb{E}\langle Z, Z_A \rangle = 6\sigma_0^{-1}.
\]

Theorem \( \bullet \) indicates that

\[
\lim_{n \to \infty} n\rho_n = \frac{3}{2}\sigma_0.
\]

In our numerical example we fix \( \sigma_0^2 = 10 \). We vary the number of data points, \( n \), ranging from 10 to 1000. For each \( n \), we use \( T = 5000 \) trials to compute empirically the optimal choice of \( \rho = \rho_n \) in order to minimize the empirical Stein loss. Furthermore, we reformulate the limiting result as

\[
\rho_n = \frac{3}{2}\sigma_0/n \iff \log(\rho_n) = -\log(n) + \log\left(\frac{3}{2}\sigma_0\right).
\]

Then, we perform a regression on \( \log(\rho_n) \) with respect to \( \log(n) \). Figure \( \bullet \) gives the relationship between \( \rho \) and \( n \) and the regression line. We can find that \( n\rho_n \) is approximately equal to a constant, which is validated by the top right plot. The plots on the left show the qualitative behavior of \( \rho_n \); the figure on the top left shows a behavior consistent with a decrease of order \( O(1/n) \), the bottom left plot shows that \( n^{1/2}\rho_n \) still decreases to zero, indicating that \( \rho_n \) converges to zero faster than the square-root rate. The regression statistics, corresponding to the regression plot shown in the bottom right of the plot, are shown in Table \( \bullet \) and \( R^2 = 0.97 \).

The theoretical constant \( \log(1.5 \cdot \sigma_0) = 1.5568 \) is very close to the empirical regression intercept 1.5525, and also the coefficient multiplying \(-\log(n)\) is close to unity. Hence, the empirical result matches perfectly with our theory.

![Figure 1: \( \rho_n \) VS \( n \) for 1-dimension normal distribution](image-url)
Table 1: Regression results for 1-dimension normal distribution

|                     | log($n$) | constant |
|---------------------|----------|----------|
| Coefficient         | -1.0037  | 1.5525   |
| 95% Confidence interval | [-1.0387,-0.9687] | [1.3419,1.7631] |

We provide additional examples involving higher dimensions. In the subsequent examples, the data is sampled from a normal distribution $N(0, \Sigma_0)$, where $(\Sigma_0)_{ij} = 10 \times 0.5^{|i-j|}, \ 1 \leq i, j \leq d$. We test the cases corresponding to $d = 3$ and $d = 5$ in the experiments. Due to computational constraints, we vary the number of data points, $n$, ranging from 20 to 400. For each $n$, we use $T = 100$ trials to compute empirically the optimal choice of uncertainty to minimize the empirical Stein loss. Figure 2 and Figure 3 show the results for 3-dimension and 5-dimension cases, respectively. Table 2 and Table 3 give the regression statistics and $R^2 = 0.97$ in both cases, and the performance is completely analogous to the one dimensional case, therefore empirically validating our theoretical results.

Table 2: Regression results for 3-dimension normal distribution

|                     | log($n$) | constant |
|---------------------|----------|----------|
| Coefficient         | -1.0340  | 2.7305   |
| 95% Confidence interval | [-1.1163,-0.9516] | [2.3045, 3.1565] |

Figure 2: $\rho_n$ VS $n$ for 3-dimension normal distribution
Figure 3: $\rho_n$ VS $n$ for 5-dimension normal distribution

| log(n)         | constant |
|---------------|----------|
| Coefficient   | -0.9177  | 2.7413   |
| 95% Confidence interval | [-0.9716, -0.8638] | [2.4625, 3.0201] |

Table 3: Regression results for 5-dimension normal distribution

References

[1] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis (3rd ed.)*. Wiley Interscience, 2003.

[2] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications, Second Edition*. Springer, 2009.

[3] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2010.

[4] Gene H Golub and Charles F Van Loan. *Matrix Computations*, volume 3. JHU Press, 2012.

[5] Robert V Hogg and Allen T Craig. *Introduction to Mathematical Statistics*. Macmillan, 1978.

[6] Erich L Lehmann and George Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2006.
[7] Viet Anh Nguyen, Daniel Kuhn, and Peyman Mohajerin Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. arXiv preprint arXiv:1805.07194, 2018.

[8] Thomas Rippl, Axel Munk, and Anja Sturm. Limit laws of the empirical Wasserstein distance: Gaussian distributions. Journal of Multivariate Analysis, 151:90–109, 2016.

[9] Cedric Villani. Topics in Optimal Transportation. Graduate Studies in Mathematics 58. American Mathematical Society, 2003.
5 Appendix: Proofs of Auxiliary Results

5.1 Proof of lemma 1

Theorem 2 (Theorem 3.1 in [7]). If $\rho > 0$ and $\hat{\Sigma}_n$ admits the spectral decomposition $\hat{\Sigma}_n = \sum_{i=1}^{d} \hat{\lambda}_i \hat{v}_i (\hat{v}_i)^T$ with eigenvalues $\hat{\lambda}_i$ and corresponding orthonormal eigenvectors $\hat{v}_i$, $i \leq d$, then the unique minimizer of (7) is given by $X^*_n(\rho) = \sum_{i=1}^{d} \hat{x}_i^* \hat{v}_i (\hat{v}_i)^T$, where

$$\hat{x}_i^* = \gamma^* \left[ 1 - \frac{1}{2} \left( \sqrt{\hat{\lambda}_i (\gamma^*)^2 + 4 \hat{\lambda}_i \gamma^* - \hat{\lambda}_i \gamma^*} \right) \right],$$

and $\gamma^* > 0$ is the unique positive solution of the algebraic equation

$$\left( \rho^2 - \frac{1}{2} \sum_{i=1}^{d} \frac{\hat{\lambda}_i}{\gamma^*} \right) \gamma - d + \frac{1}{2} \sum_{i=1}^{d} \sqrt{\hat{\lambda}_i (\gamma^*)^2 + 4 \hat{\lambda}_i \gamma} = 0. \quad (8)$$

Since the underlying covariance matrix is invertible then, if $n > d$, with probability 1, we have $\hat{\lambda}_i > 0$ for $i = 1, 2, \ldots, d$.

When $\gamma \rightarrow \infty$ and $\hat{\lambda}_i > 0$, we have following equation,

$$\sqrt{\hat{\lambda}_i^2 \gamma^2 + 4 \hat{\lambda}_i \gamma} - (\hat{\lambda}_i \gamma + 2) = -\frac{4}{\sqrt{\hat{\lambda}_i^2 \gamma^2 + 4 \hat{\lambda}_i \gamma} + (\hat{\lambda}_i \gamma + 2)} = -\frac{2}{\hat{\lambda}_i \gamma} + O \left( \frac{1}{\gamma^2} \right)$$

and (8) becomes

$$\left( \rho^2 - \frac{1}{2} \sum_{i=1}^{d} \hat{\lambda}_i \right) \gamma - d + \frac{1}{2} \sum_{i=1}^{d} \left( \hat{\lambda}_i \gamma^* + 2 \right) - \frac{2}{\hat{\lambda}_i \gamma^*} + O \left( \frac{1}{(\gamma^*)^2} \right) = 0.$$ 

After simplification, we have

$$\rho^2 \gamma^* = \sum_{i=1}^{d} \frac{1}{\hat{\lambda}_i \gamma^*} + O \left( \frac{1}{(\gamma^*)^2} \right).$$

Therefore, we solve for

$$\frac{1}{\gamma^*} = \frac{1}{\sqrt{\sum_{i=1}^{d} \hat{\lambda}_i^{-1}}} \left( \rho + O(\rho^2) \right). \quad (9)$$

By plugging it to (8), we have

$$\hat{x}_i^* = \frac{2 \gamma^*}{\sqrt{\hat{\lambda}_i^2 (\gamma^*)^2 + 4 \hat{\lambda}_i \gamma^* + \hat{\lambda}_i \gamma^* + 2}} = \frac{1}{\hat{\lambda}_i + 2/\gamma^*} + O \left( \frac{1}{(\gamma^*)^2} \right).$$
Then, from (9), we have
\[
\frac{1}{\lambda_i + 2/\gamma^*} = \frac{1}{\lambda_i + \sqrt{\frac{2(\rho + O(\rho^2))}{\sum_{i=1}^{\infty} \lambda_i}}} = \frac{1}{\lambda_i} - \frac{2\rho}{\lambda_i \sqrt{\sum_{i=1}^{d} \lambda_i}} + O(\rho^2).
\]

Therefore, we conclude
\[
x_i^* = \frac{1}{\lambda_i} - \frac{2\rho}{\lambda_i^2 \sqrt{\sum_{i=1}^{d} \lambda_i}} + O(\rho^2),
\]
and
\[
X_n^*(\rho) = \hat{\Sigma}_n^{-1} + \hat{A}_n \rho + O(\rho^2),
\]
where
\[
\hat{A}_n = -\sum_{i=1}^{d} \frac{2\hat{v}_i (\hat{v}_i)^T}{\lambda_i^2 \sqrt{\sum_{i=1}^{d} \lambda_i}} = -\frac{2}{\sqrt{\text{tr}(\hat{\Sigma}_n^{-1})}} \Sigma^{-2}.
\]
Furthermore, since
\[
\left. \frac{\partial X_n^*(\rho)^{-1}}{\partial \rho} \right|_{\rho=0} = -\hat{\Sigma}_n \hat{A}_n \hat{\Sigma}_n,
\]
we have
\[
X_n^*(\rho)^{-1} = \hat{\Sigma}_n - \hat{\Sigma}_n \hat{A}_n \hat{\Sigma}_n \rho + O(\rho^2).
\]

5.2 Proof of Proposition

Proof of (2): \(\hat{\Sigma}_n\) is the average of i.i.d copies \(\xi_i \xi_i^T\), then the result follows by CLT.

Proof of (3): let \(f(\Sigma) = -2 (\text{tr}(\Sigma^{-1}))^{-1/2} \Sigma^{-2}\), where \(\Sigma\) is positive-definite matrix. We now expand \(f(\Sigma + hA)\) for any matrix \(A\) as the scalar \(h > 0\) tends to zero to obtain an representation for the gradient of \(f(\Sigma)\), \(Df(\Sigma)\), this expansion yields
\[
f(\Sigma + hA) = -2 (\text{tr}((\Sigma + hA)^{-1}))^{-1/2} (\Sigma + hA)^{-2}
\]
\[
= -2 (\text{tr}(\Sigma^{-1}) - \text{tr}(h\Sigma^{-1}A\Sigma^{-1}) + o(h))^{-1/2} \left( (I + h\Sigma^{-1}A)^{-1} \Sigma^{-1} \right)^2
\]
\[
= -2 \left( \text{tr}(\Sigma^{-1})^{-1/2} \left( 1 - h \frac{\text{tr}(\Sigma^{-1}A\Sigma^{-1})}{\text{tr}(\Sigma^{-1})} \right) \left( \Sigma^{-2} - h\Sigma^{-1}A\Sigma^{-2} - h\Sigma^{-2}A\Sigma^{-1} \right) + o(h) \right)^{-1/2}
\]
\[
= -2 \left( \text{tr}(\Sigma^{-1})^{-1/2} \left( 1 + h \frac{\text{tr}(\Sigma^{-1}A\Sigma^{-1})}{\text{tr}(\Sigma^{-1})} \right) \left( \Sigma^{-2} - h\Sigma^{-1}A\Sigma^{-2} - h\Sigma^{-2}A\Sigma^{-1} \right) + o(h) \right)^{-1/2}
\]
\[
= -2 \left( \text{tr}(\Sigma^{-1})^{-1/2} \left( \Sigma^{-2} + h \frac{\text{tr}(\Sigma^{-1}A\Sigma^{-1})}{\text{tr}(\Sigma^{-1})} \Sigma^{-2} \right) - h\Sigma^{-1}A\Sigma^{-2} - h\Sigma^{-2}A\Sigma^{-1} + o(h) \right)^{-1/2}
\]
\[
= f(\Sigma) - h \frac{\text{tr}(\Sigma^{-1}A\Sigma^{-1})}{\text{tr}(\Sigma^{-1})^{3/2}} \Sigma^{-2} + 2h \frac{\Sigma^{-1}A\Sigma^{-2} + \Sigma^{-2}A\Sigma^{-1}}{\text{tr}(\Sigma^{-1})^{1/2}} + o(h),
\]

11
which, in turn, results in the linear operator satisfying for any $A \in \mathbb{R}^{d \times d}$

$$Df(\Sigma)A = -\frac{\text{tr} \left( (\Sigma^{-1}A) \Sigma^{-2} \right)}{\text{tr}(\Sigma^{-1})^{3/2}} + 2 \frac{\text{tr} \left( A \Sigma^{-2} + \Sigma^{-2} A \Sigma^{-1} \right)}{\text{tr}(\Sigma^{-1})^{1/2}}. \quad (10)$$

After applying delta method, we have the desired result.

### 5.3 Proof of Proposition 2

We first note the following elementary result, which is standard in matrix algebra (see, for example, Chapter 2.3 in [4]).

**Lemma 3.** For any $d \times d$ matrices $A, B$ (real valued) we have

$$\text{tr}(A^T A) \text{tr}(B^T B) \geq \text{tr}(A^T B)^2 = |\langle A, B \rangle|^2$$

strict inequality holds unless $A$ is a multiple of $B$.

Now we proceed with the proof of Proposition 2. It suffices to show that $\langle Z, Z_A \rangle \geq 0$ with probability one and that $\langle Z, Z_A \rangle > 0$ with positive probability. Note that

$$\langle Z, Z_A \rangle = -\frac{\text{tr}(Z \Sigma^{-2})^2}{\text{tr}(\Sigma^{-1})^{3/2}} + 2 \frac{\text{tr} \left( \Sigma^{-1} Z \Sigma^{-2} Z \right)}{\text{tr}(\Sigma^{-1})^{1/2}}.$$

We claim that

$$\text{tr}(\Sigma^{-1}) \text{tr} \left( \Sigma^{-1} Z \Sigma^{-2} Z \right) \geq \text{tr}(Z \Sigma^{-2})^2,$$

follows from Lemma 3, this will imply that $\langle Z, Z_A \rangle > 0$. We use the Polar factorization (see, for example, Chapter 4.2 in [4]) for positive definite matrices. That is, we write $\Sigma^{1/2} \Sigma^{1/2} = \Sigma_0$, where $\Sigma_0^{1/2}$ is a symmetric positive definite matrix. Note that we can write

$$Z = \Sigma_0^{1/2} W \Sigma_0^{1/2},$$

where $W = \Sigma_0^{-1/2} Z \Sigma_0^{-1/2}$ is a symmetric matrix. To recover the matrices $A$ and $B$, we let

$$A = \Sigma_0^{-1/2}, \quad S = \Sigma_0 \quad \text{and} \quad B = WS^{-1/2}.$$

Note that

$$= \text{tr} \left( \Sigma_0^{-1} Z \Sigma_0^{-2} Z \right) = \text{tr} \left( Z \Sigma_0^{-1} \cdot \Sigma_0^{-1} Z \Sigma_0^{-1} \right)$$

$$= \text{tr} \left( S^{1/2} W S^{1/2} S^{-1/2} S^{-1/2} S^{-1/2} \cdot S^{-1/2} S^{-1/2} \left( S^{1/2} W S^{1/2} \right) S^{-1/2} S^{-1/2} \right)$$

$$= \text{tr} \left( S^{1/2} W S^{-1/2} \cdot S^{-1/2} W S^{-1/2} \right) = \text{tr} \left( S^{-1/2} W S^{-1/2} \right) = \text{tr} \left( (WS^{-1/2})^T S^{-1/2} W \right).$$

So, this verifies that the choice of $B$ is consistent with the use of Lemma 3. Clearly, $AA^T = \Sigma_0^{-1}.$
so this choice is also consistent with Lemma 3. Finally, we have that
\[
\text{tr}(Z\Sigma_0^{-2}) = \text{tr}(S^{1/2}WS^{1/2}S^{-1/2}S^{-1/2}\Sigma_0^{-1}) = \text{tr}(S^{1/2}WS^{-1/2}S^{-1/2}) = \text{tr}(WS^{-1/2}S^{-1/2}) = \text{tr}(S^{-1/2}S^{-1/2}W) = \text{tr}(A^TB).
\]
The result then follows.

5.4 Proof of Lemma 2

We first collect a few results from linear algebra (see, for example, Chapter 2.3 in [4]).

**Lemma 4.** For any \(d \times d\) matrix \(A\) (real valued) we define \(\|A\|_F^2 = \langle A, A \rangle = \text{tr}(A^TA)\) (the Frobenius norm) and let \(\|A\|_2^2 = |\lambda_{\text{max}}(A^TA)|\) (where \(\lambda_{\text{max}}(B)\) is the eigenvalue of largest modulus of the matrix \(B\)). Then, for any \(A, B\) matrices of size \(d \times d\) with real valued elements we have
\[
\|AB\|_F \leq \|A\|_2 \|B\|_F, \quad \|B\|_2 \leq \|B\|_F.
\]

In addition, we have the following properties of the distribution of \(\hat{\Sigma}_n\), which follows the Wishart law (see, for example, Chapter 7 in [1]).

**Lemma 5.** Let us write \(\xi_i = C\zeta_i\) where \(CC^T = \Sigma_0\) and put
\[
S_n = C \left( \sum_{i=1}^n \xi_i^T \xi_i \right) C^T,
\]

note that \(\hat{\Sigma}_n = S_n/n\). Then, \(S_n\) is distributed Wishart with parameters \(d, n\) and \(\Sigma_0\) (denoted \(W_d(n, \Sigma_0)\)). Equivalently, \(W = C^{-1}S_n(C^T)^{-1}\) is distributed \(W_d(n, I)\). Moreover, the eigenvalue distribution of \(W\) satisfies
\[
f_{w_1, \ldots, w_d} (w_1, \ldots, w_d) = c_d \prod_{i=1}^d \frac{\exp(-w_i/2)}{2^{n/2}\Gamma((n-i+1)/2)} \frac{w_i^{(n-d+1)/2-1}}{\prod_{j>i} (w_j - w_i) \ I(0 < w_1 < \ldots < w_d)},
\]

where \(c_d\) is a constant independent of \(n\).

Now we are ready to provide the proof of Lemma 2. First, from Lemma 5 we have that
\[
f_{w_1, \ldots, w_d} (w_1, \ldots, w_d) = c_d \prod_{i=1}^d \frac{\exp(-w_i/2)}{2^{n/2}\Gamma((n-i+1)/2)} \frac{w_i^{(n-d+1)/2-1}}{\prod_{j>i} (w_j - w_i) \ I(0 < w_1 < \ldots < w_d)}.
\]
We know that \( \Lambda_i \rightarrow \infty \) as \( n \rightarrow \infty \). It suffices to show that for any \( \alpha \rightarrow \infty \) the previous identity can be interpreted as follows. Let \( W(n) := (W_{(1)}^{(n)}, \ldots, W_{(d)}^{(n)}) \) be the eigenvalues of a \( W_d(n, I) \) random matrix and let \( \Lambda(n) := (\Lambda_1(n), \ldots, \Lambda_d(n)) \) be independent random variables such that \( \Lambda_i(n) \sim \chi^2_{n-i+1} \), then for any positive (and measurable) function \( g : \mathbb{R}^d \rightarrow [0, \infty) \) we have that

\[
E \left[ g \left( W(n) \right) \right] \leq c_d \frac{d}{d} E \left[ g \left( \Lambda(n) \right) \prod_{i=1}^{d} \frac{\Lambda_i(n)}{n}^{(d-i)/2} \prod_{j>i} n^{1/2} \left( \Lambda_j(n) / \Lambda_i(n) - 1 \right) \right] . \tag{12}
\]

To verify the first statement of Lemma 3 we need to show the uniform integrability of \( \left\langle \hat{\Sigma}_n \hat{A}_n \hat{\Sigma}_n, \hat{A}_n \right\rangle \). In turn, it suffices to verify that for some \( r > 1 \) and some \( n_0 < \infty \) we have that

\[
\sup_{n \geq n_0} E \left[ \left| \left\langle \hat{\Sigma}_n \hat{A}_n \hat{\Sigma}_n, \hat{A}_n \right\rangle \right|^r \right] < \infty ,
\]

see, for example, Chapter 5 in [3].

Applying Lemma 3 and Lemma 4 together with (12) and repeated use of Hölder’s inequality it suffices to show that for any \( r > 1 \) there exists \( n_0 \) such that

\[
\sup_{n \geq n_0} \left( E \left[ \left( \frac{\Lambda_i}{n} \right)^r \right] \cdot E \left[ \left( \frac{n}{\Lambda_i} \right)^r \right] \cdot E \left[ \left| n^{1/2} \left( \Lambda_j(n) / \Lambda_i(n) - 1 \right) \right|^r \right] \right) < \infty . \tag{13}
\]

We know that \( \Lambda_i(n)/n \) follows a Gamma distribution with shape parameter \( \alpha = (n - i + 1)/2 \) and scale parameter \( \lambda = n/2 \). Write \( Y_n \sim Gamma(\alpha, \lambda) \) and note that

\[
E \left( \frac{1}{Y_n^r} \right) = \int_0^\infty \frac{1}{y^r} \frac{\exp(-\lambda y) \lambda^\alpha y^{\alpha-1}}{\Gamma(\alpha)} dy = \frac{\Gamma(\alpha - r) \lambda^r}{\Gamma(\alpha)} . \tag{14}
\]

It follows from standard properties of the Gamma function that \( n^r \Gamma(\alpha - r) \lambda^r / \Gamma(\alpha) = O(1) \) as \( \alpha \rightarrow \infty \) (see, for example, Chapter 3 in [3]). Exactly the same approach can be used to study
\[ \mathbb{E} \left[ (\Lambda_j/n)^r \right], \text{thus concluding the first part of Lemma 2} \]

Now, note that
\[
\mathbb{E} \left[ n^{1/2} (\Lambda_j/n/\Lambda_i(n) - 1)^r \right] = \mathbb{E} \left[ n^{1/2} (\Lambda_j/n/\Lambda_i(n) - 1)^r \right] \mathbb{I} (|\Lambda_j(n)/n - 1| \leq \varepsilon, |\Lambda_i(n)/n - 1| \leq \varepsilon) + \mathbb{E} \left[ n^{1/2} (\Lambda_j/n/\Lambda_i(n) - 1)^r \right] \mathbb{I} (|\Lambda_j(n)/n - 1| > \varepsilon \cup |\Lambda_i(n)/n - 1| > \varepsilon) .
\]

It is straightforward to verify (for example by computing moment generating functions) that
\[
\sup_{n \geq 1} \mathbb{E} \left( n^{r/2} \left| \frac{\Lambda_j(n)}{n} - 1 \right|^r \right) < \infty \tag{15}
\]
for any \( r > 0 \) and further, we can conclude that
\[
\sup_{n \geq 1} \mathbb{E} \left[ n^{1/2} (\Lambda_j/n/\Lambda_i(n) - 1)^r \right] I(|\Lambda_j(n)/n - 1| \leq \varepsilon, |\Lambda_i(n)/n - 1| \leq \varepsilon) \leq \frac{2^{r-1}}{(1-\varepsilon)^r} \sup_{n \geq 1} \mathbb{E} \left( n^{r/2} \left| \frac{\Lambda_j(n)}{n} - 1 \right|^r \right) < \infty.
\]

Then, because \( \Lambda_j(n)/n \) (being the sum of \( n - j + 1 \) i.i.d. random variables with finite moment generating function) satisfies a large deviations principle (see, for instance, Chapter 2.2 in [2]), we have for \( s, t \in (0, \infty) \) with \( 1/s + 1/t = 1 \) that
\[
\mathbb{E} \left[ n^{1/2} (\Lambda_j(n)/\Lambda_i(n) - 1)^r \mathbb{I} (|\Lambda_j(n)/n - 1| > \varepsilon) \right] \leq \mathbb{E}^{1/s} \left[ n^{1/2} (\Lambda_j(n)/\Lambda_i(n) - 1)^{s^r} \right] \mathbb{P}^{1/t} (|\Lambda_j(n)/n - 1| > \varepsilon) . \tag{16}
\]

Because of our discussion involving the finiteness of the first two factors in (13) we conclude that the first term in the right hand side of (16) grows at rate \( O(n^{r/2}) \), which is polynomial, whereas the second term, due to the large deviations principle invoked earlier converges exponentially fast to zero for each \( \varepsilon > 0 \). Therefore, we conclude (15).

For the second part of Lemma 2, note that Lemma 3 implies
\[
\left| \left\langle \sqrt{n} \left( \Sigma_n - \Sigma_0 \right), \hat{A}_n \right\rangle \right|^2 \leq \left\| \sqrt{n} \left( \Sigma_n - \Sigma_0 \right) \right\|_F^2 \left\| \hat{A}_n \right\|_F^2 .
\]

Then, we directly have the uniform integrability of \( \left\| \sqrt{n} \left( \Sigma_n - \Sigma_0 \right) \right\|_F^2 \) and from the earlier bounds leading to the analysis of (14) we conclude the uniform integrability of \( \left\| \hat{A}_n \right\|_F^2 \). So, the second part of Lemma Lemma 2 follows.
Finally, for the third part of Lemma 2 let us write 
\[ \hat{A}_n = g \left( \hat{\Sigma}_n \right), \]
where
\[ g(\Sigma) = -2 \sqrt{\text{tr}(\Sigma^{-1})} \Sigma^{-2}. \]

The argument is similar to that given to establish (15). We have argued that \( f(\cdot) \) is smooth around \( \Sigma_0 \) (this was the basis for the use of the delta method earlier in our argument). Moreover, note that \( \hat{\Sigma}_n \) satisfies a large deviations principle. Therefore
\[ \left| \left| \left( \sqrt{n} (\hat{\Sigma}_n - \Sigma_0), \sqrt{n} (\hat{A}_n - A_0) \right) \right| \right|_F \leq \varepsilon. \]

By applying Lemma 3 and the fact that \( Df(\cdot) \) is continuous around \( \Sigma_0 \) (see the expression of \( Df(\cdot) \) in (10)) we conclude that
\[ \left| \left| \left( \sqrt{n} (\hat{\Sigma}_n - \Sigma_0), \sqrt{n} (\hat{g}(\hat{\Sigma}_n) - A_0) \right) \right| \right|_F \leq \varepsilon. \]

The right hand side is seen to be uniformly integrable by standard properties of the Gaussian distribution. On the other hand, we have that
\[ \mathbb{E} \left( \left| \left| \left( \sqrt{n} (\hat{\Sigma}_n - \Sigma_0), \sqrt{n} (\hat{g}(\hat{\Sigma}_n) - A_0) \right) \right| \right|_F > \varepsilon \right) \]
\[ \leq \mathbb{E}^{1/2} \left( \left| \left| \left( \sqrt{n} (\hat{\Sigma}_n - \Sigma_0), \sqrt{n} (\hat{g}(\hat{\Sigma}_n) - A_0) \right) \right| \right|_F \right)^2 \cdot \mathbb{P} \left( \left| \left| \hat{\Sigma}_n - \Sigma_0 \right| \right|_F > \varepsilon \right)^{1/2}. \]

We have argued throughout the proof of the first part of Lemma 2 that
\[ \mathbb{E} \left( \left| \left| \sqrt{n} (\hat{g}(\hat{\Sigma}_n) - A_0) \right| \right|_F \right) = O \left( n^{r/2} \right) \]
but, on the other hand, \( \mathbb{P} \left( \left| \left| \hat{\Sigma}_n - \Sigma_0 \right| \right|_F > \varepsilon \right) = O \left( \exp \left( -cn \right) \right) \) for some \( c > 0. \) Therefore, using Lemma 3 and Lemma 4 and the previous estimates we can conclude the last part of Lemma 2.