Optimal rates of convergence for covariance matrix estimation

Alberto Bordino

July 17, 2023

1 Introduction

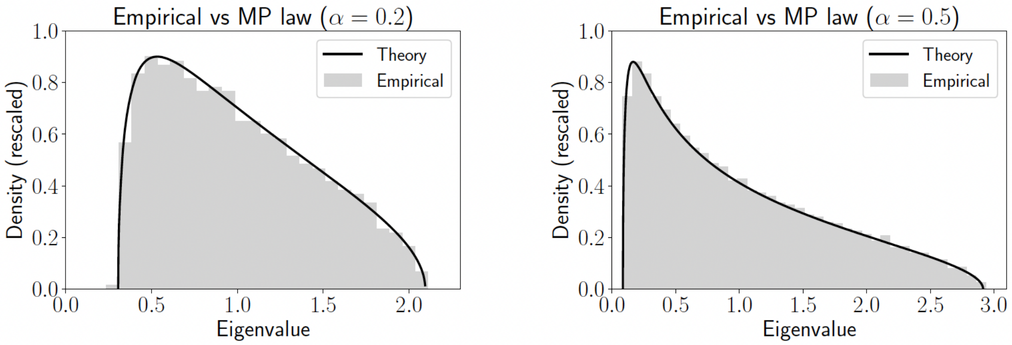
Suppose we observe independent and identically distributed p-variate random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with covariance matrix $\Sigma_{p\times p}$ and the goal is to estimate the unknown matrix $\Sigma_{p\times p}$ based on the sample $\{\mathbf{X}_i : i = 1, \ldots, n\}$. The most natural choice would be to use the sample covariance matrix

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X_i} - \bar{\mathbf{X}})(\mathbf{X_i} - \bar{\mathbf{X}})^T,$$

(or the bias corrected version with $1/n-1$ in place of $1/n$) but it was shown to perform poorly in high dimensional settings, i.e. when $p/n \to \alpha \in (0,1)$. To convince ourselves that this is actually the case, suppose that $d/n = \alpha \in (0,1)$ and $\Sigma = I_d$, with each sample $X_i \sim N(0, I_d)$ for $i = 1, \ldots, n$. Using these n samples, we generated the sample covariance matrix, and then computed its vector of eigenvalues $\gamma(\widehat{\Sigma}) \in \mathbb{R}^d$, say arranged in non-increasing order as

$$\gamma_{\max}(\widehat{\Sigma}) = \gamma_1(\widehat{\Sigma}) \geq \gamma_2(\widehat{\Sigma}) \geq \cdots \geq \gamma_d(\widehat{\Sigma}) = \gamma_{\min}(\widehat{\Sigma}) \geq 0$$

The plots below show a histogram of the vector $\gamma(\widehat{\Sigma}) \in \mathbb{R}^d$ of eigenvalues: Figure 1(a) corresponds to the case $(n, d) = (4000, 800)$ or $\alpha = 0.2$, whereas Figure 1(b) shows the pair $(n, d) = (4000, 2000)$ or $\alpha = 0.5$. If the sample covariance matrix were converging to the identity matrix, then the vector of eigenvalues $\gamma(\widehat{\Sigma})$ should converge to the all-ones vector, and the corresponding histograms should concentrate around 1. Instead, the histograms in both plots are highly dispersed around 1, with differing shapes depending on the aspect ratios. These shapes are characterized by an asymptotic distribution known as the Marčenko-Pastur law. Apart from this, the key message here is to realize that the sample covariance estimator is a consistent estimator of Σ when $d/n \to 0$, whereas if



(a)

(b)

Figure 1: Simulations showing the Marčenko-Pastur law.

$d/n = \alpha \in (0,1)$ this is no more the case!

Can we do better in high dimensions? The idea of Cai et al. (2010) is to use regularisation and shrink down the off-diagonal entries of the sample covariance matrix progressively to zero to get a new estimator, which they call tapering estimator. This is a similar approach to Bickel and Levina (2008), where they introduce the so-called banding estimator, whose entries are also shrunk to zero, but not progressively. The choice of shrinking the entries of the sample covariance matrix down to zero linearly, and not abruptly, makes the all difference since the tapering estimator of Cai et al. (2010) has been proven to be minimax optimal with respect to both the spectral and the Frobenius norms over the class of distributions with covariance matrix in $\mathcal{F}_\alpha$ defined below, whereas the banding estimator of Bickel and Levina (2008) is not.

In what follows, we will define the tapering estimator formally, and focus on its risk with respect to the spectral norm only. In particular, we will state Theorem 2 of Cai et al. (2010), which gives an upper bound for the risk, and Theorem 3, which shows that this estimator is minimax optimal. Before moving on, we conclude this introduction with some notation. We will denote with $\mathbf{X} \sim SG(\rho)$ a sub-Gaussian random vector with proxy ρ^2, and with $||A|| = \max\{\sigma_{max}(A), |\sigma_{min}(A)|\}$ the spectral norm of the matrix A, where $\sigma_{max}, \sigma_{min}$ are the maximum and minimum singular value respectively. Finally, we indicate with $\mathcal{F}_\alpha$ the class of matrices

$$\mathcal{F}_\alpha = \mathcal{F}_\alpha\left(M_0, M\right) = \left\{\Sigma : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \leq Mk^{-\alpha} \text{ for all } k, \text{ and } \lambda_{\max}(\Sigma) \leq M_0\right\},$$

where $\lambda_{\max}(\Sigma)$ is the maximum eigenvalue of the matrix Σ, and $\alpha > 0, M > 0$, where α can be interpreted as the smoothing parameter in nonparametric function estimation problems.

2 The tapering estimator

Let be σ^*_{ij} the $(i,j)-$th entry of the sample covariance matrix S_n. For a given even integer k with $1 \leq k \leq p$, we define a tapering estimator as

$$\widehat{\Sigma} = \widehat{\Sigma}_k = \left(w_{ij}\sigma^*_{ij}\right)_{p\times p},$$

and the weights

$$w_{ij} = k_h^{-1}\left\{(k - |i-j|)_+ - (k_h - |i-j|)_+\right\},$$

where $k_h = k/2$. Without loss of generality, we assume that k is even. Note that the weights w_{ij} can be rewritten as

$$w_{ij} = \begin{cases} 1, & \text{when } |i-j| \leq k_h, \\ 2 - \frac{|i-j|}{k_h}, & \text{when } k_h < |i-j| < k, \\ 0, & \text{otherwise.} \end{cases}$$

Here, $w = (w_{ij})_{p\times p}$ is a matrix of weights which governs the decay towards zero of the σ^*_{ij} based on the integer k, which a tuning parameter that can be chosen to get the sharpest upper bound possible. An illustration of the tapering estimator is given in Figure 2.

Now, define by $\mathcal{P}_\alpha(M_0, M, \rho)$ the class of probability distributions such that $\mathbf{X} \sim SG(\rho)$ with covariance matrix in $\mathcal{F}_\alpha(M_0, M)$. Then we have the following result:

Theorem 1 (Theorem 2 in Cai et al. (2010)). *The tapering estimator $\widehat{\Sigma}_k$ of the covariance matrix $\Sigma_{p\times p}$ with*

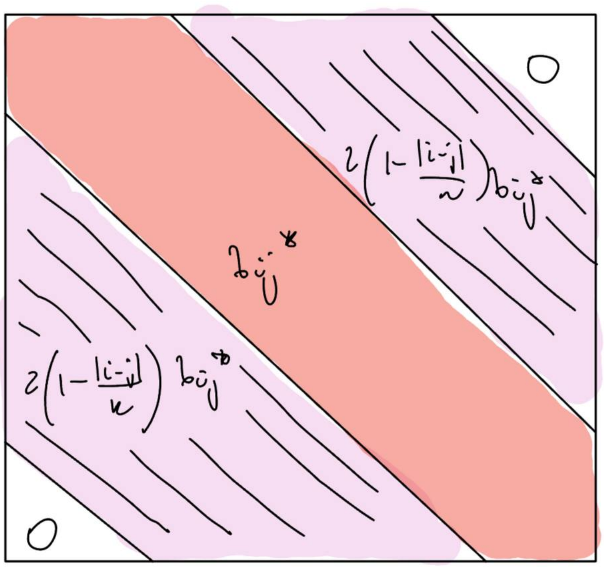


Figure 2: Tapering estimator. The entries are shrunk down towards zero in a linear fashion depending on the weights $(w_{ij})_{p\times p}$.

$p \geq n^{1/(2\alpha+1)}$ *satisfies*

$$\sup_{\mathcal{P}_\alpha} \mathbb{E}\left\|\widehat{\Sigma}_k - \Sigma\right\|^2 \leq C\frac{k+\log p}{n} + Ck^{-2\alpha}$$

for $k = o(n), \log p = o(n)$ *and some constant* $C > 0$. *In particular, the estimator* $\widehat{\Sigma} = \widehat{\Sigma}_k$ *with* $k = n^{1/(2\alpha+1)}$ *satisfies*

$$\sup_{\mathcal{P}_\alpha} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|^2 \leq Cn^{-2\alpha/(2\alpha+1)} + C\frac{\log p}{n}.$$

It is clear that the optimal choice of k is of order $n^{1/(2\alpha+1)}$. The upper bound is thus rate optimal among the class of the tapering estimators, but, as the next result shows, the estimator $\widehat{\Sigma}_k$ with $k = n^{1/(2\alpha+1)}$ is in fact rate optimal among all estimators.

Theorem 2 (Theorem 3 in Cai et al. (2010)). *Suppose* $p \leq \exp(\gamma n)$ *for some constant* $\gamma > 0$. *The minimax risk for estimating the covariance matrix* Σ *over* $\mathcal{P}_\alpha$ *under the operator norm satisfies*

$$\inf_{\widehat{\Sigma}} \sup_{\mathcal{P}_\alpha} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|^2 \geq cn^{-2\alpha/(2\alpha+1)} + c\frac{\log p}{n}.$$

The proof of this result is based on the construction of a finite collection of multivariate normal distributions, for which we calculate the total variation affinity between pairs of probability measures in the collection. Before moving to this, it is worth mentioning as a final remark that the optimal rate is of the order

$$n^{-2\alpha/(2\alpha+1)} + \frac{\log p}{n},$$

which is the sum of a parametric and a nonparametric rate. The latter is of the same order as the optimal rate for nonparametric density estimation with respect to the L^2 norm over the class of α-Holder continuous functions. The similarity is simple: the level of smoothness here is represented by how fast the off-diagonal entries of Σ goes to zero, and the faster they decay, the better our estimator. Moreover, if p is small, say $p = o\left(n^{1/(2\alpha+1)}\right)$, p has no effect on the convergence rate and the rate is purely driven by the smoothness parameter α. However, when p is large, that is, $\log p \gg n^{1/(2\alpha+1)}$, p plays a significant role in determining the minimax rate. Furthermore, coming back to the banding estimator

$$\widehat{\Sigma}_B = \left(\sigma^*_{ij} I\{|i-j| \leq k\}\right)$$

given in Bickel and Levina (2008), it is easy to see that is not rate optimal. The thigher upper bound is attained when

$$k = \left(\frac{\log p}{n}\right)^{1/(2(\alpha+1))},$$

leading to the rate of convergence

$$\left(\frac{\log p}{n}\right)^{\alpha/(\alpha+1)}.$$

But if you take, for example, $\alpha = 1/2$ and $p = e^{\sqrt{n}}$, Bickel and Levina (2008)'s rate is $n^{-1/6}$, while the optimal rate given in Cai et al. (2010) is $n^{-1/2}$.

3 Proof of minimax optimality

As stated above, the first step of the proof is to define a suitable class of multivariate normal distributions. To this aim, for given positive integers k and m with $2k \leq p$ and $1 \leq m \leq k$, define the $p \times p$ matrix $B(m,k) = (b_{ij})_{p\times p}$ with

$$b_{ij} = I\{i = m \text{ and } m+1 \leq j \leq 2k, \text{ or } j = m \text{ and } m+1 \leq i \leq 2k\}.$$

Set $k = n^{1/(2\alpha+1)}$ and $a = k^{-(\alpha+1)}$. The generic matrix $B(m,k)$ is shown in Figure 3.

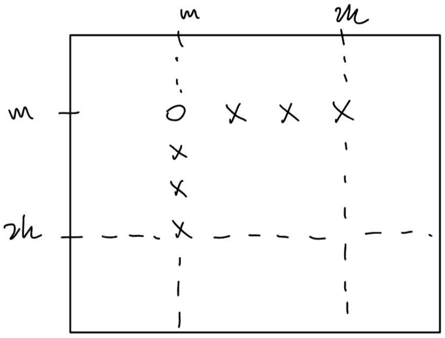


Figure 3: Visual representation of the matrix $B(m,k)$.

We then define the collection of 2^k covariance matrices as

$$\mathcal{F}_{11} = \left\{\Sigma(\theta) : \Sigma(\theta) = I_p + \tau a \sum_{m=1}^{k} \theta_m B(m,k), \theta = (\theta_m) \in \{0,1\}^k\right\},$$

where I_p is the $p \times p$ identity matrix and $0 < \tau < 2^{-\alpha-1}M$. Without loss of generality, we assume that $M_0 > 1$ and $\rho > 1$. Otherwise, we can replace I_p by εI_p for $0 < \varepsilon < \min\{M_0, \rho\}$. For $0 < \tau < 2^{-\alpha-1}M$ it is easy to check that $\mathcal{F}_{11} \subset \mathcal{F}_\alpha(M_0, M)$ as $n \to \infty$. In addition to $\mathcal{F}_{11}$, we also define a collection of diagonal matrices

$$\mathcal{F}_{12} = \left\{\Sigma_m : \Sigma_m = I_p + \left(\sqrt{\frac{\tau}{n}\log p_1} I\{i = j = m\}\right)_{p\times p}, 0 \leq m \leq p_1\right\},$$

where $p_1 = \min\left\{p, e^{n/2}\right\}$ and $0 < \tau < \min\left\{(M_0-1)^2, (\rho-1)^2, 1\right\}$. Let $\mathcal{F}_1 = \mathcal{F}_{11} \cup \mathcal{F}_{12}$. It is clear that $\mathcal{F}_1 \subset \mathcal{F}_\alpha(M_0, M)$. Then, the strategy is as follows: we show that

1. $\inf_{\widehat{\Sigma}} \sup_{\mathcal{F}_{11}} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|^2 \geq cn^{-2\alpha/(2\alpha+1)}$ using Assouad's lemma
2. $\inf_{\widehat{\Sigma}} \sup_{\mathcal{F}_{12}} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|^2 \geq c\log p/n$ using Le Cam's method,

and this is enough for proving the lower bound since

$$\inf_{\widehat{\Sigma}} \sup_{\mathcal{F}_\alpha} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|^2 \geq \inf_{\widehat{\Sigma}} \sup_{\mathcal{F}_1} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|^2 = \inf_{\widehat{\Sigma}} \sup_{\mathcal{F}_{11} \cup \mathcal{F}_{12}} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|^2 \geq$$
$$\frac{1}{2}\left(\inf_{\widehat{\Sigma}} \sup_{\mathcal{F}_{11}} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|^2 + \inf_{\widehat{\Sigma}} \sup_{\mathcal{F}_{12}} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|^2\right) \geq \frac{c}{2}\left(n^{-2\alpha/(2\alpha+1)} + \frac{\log p}{n}\right).$$

3.1 Proof of $\inf_{\widehat{\Sigma}} \sup_{\mathcal{F}_{11}} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|^2 \geq cn^{-2\alpha/(2\alpha+1)}$ using Assouad's lemma

Assouad's lemma can be seen as a generalisation of Le Cam's method where we consider multiple, say k, pairwise comparisons at the same time, instead of the classical two-points testing procedure of Le Cam. As a drawback, we must point out that Assouad's method cannot be applied for certain loss functions, and can be applied only when we can upper bound the risk we are considering with the maximum risk on the hypercube of a certain dimension k. For a more detailed discussion of such limitations, refer to Tsybakov (2009). Nonetheless, if these conditions are satisfied, we are able to push up our lower bound by a factor of k corresponding to the dimensionality of the associated hypercube.

Lemma 3. *Let $\Theta = \{0,1\}^k$ and let T be an estimator based on an observation from a distribution in the collection $\{P_\theta, \theta \in \Theta\}$. Then for all $s > 0$*

$$\max_{\theta\in\Theta} 2^s \mathbb{E}_\theta d^s(T, \psi(\theta)) \geq \min_{H(\theta,\theta')\geq 1} \frac{d^s(\psi(\theta), \psi(\theta'))}{H(\theta,\theta')} \cdot \frac{k}{2} \cdot \min_{H(\theta,\theta')=1} \|\mathbb{P}_\theta \wedge \mathbb{P}_{\theta'}\|.$$

For the proof, refer to Lemma 2.12 in Tsybakov (2009), or to this beautiful blog. Assouad's lemma is connected to multiple comparisons. In total, there are k comparisons. The lower bound has three factors. The first factor is basically the minimum cost of making a mistake per comparison, and the last factor is the lower bound for the total probability of making type I and type II errors for each comparison, and $k/2$ is the expected number of mistakes one makes when $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ are not distinguishable from each other when $H(\theta,\theta') = 1$.

Consider now $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{\text{i.i.d.}}{\sim} N(0, \Sigma(\theta))$ with $\Sigma(\theta) \in \mathcal{F}_{11}$. Denote the joint distribution by P_θ. Applying Assouad's lemma to the parameter space $\mathcal{F}_{11}$, we have

$$\inf_{\widehat{\Sigma}} \max_{\theta\in\{0,1\}^k} 2^2 E_\theta \|\widehat{\Sigma} - \Sigma(\theta)\|^2$$
$$\geq \min_{H(\theta,\theta')\geq 1} \frac{\|\Sigma(\theta) - \Sigma(\theta')\|^2}{H(\theta,\theta')} \frac{k}{2} \min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\|.$$

As for the first factor, Lemma 5 in Cai et al. (2010) gives

$$\min_{H(\theta,\theta')\geq 1} \frac{\|\Sigma(\theta) - \Sigma(\theta')\|^2}{H(\theta,\theta')} \geq cka^2$$

for some universal constant $c > 0$. For the third one, Lemma 6 in Cai et al. (2010) states that, if $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{\text{i.i.d.}}{\sim}$

$N(0, \Sigma(\theta))$ with $\Sigma(\theta) \in \mathcal{F}_{11}$ and if P_θ denote the joint distribution, then

$$\min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\| \geq c$$

for some universal constant $c > 0$. Combining Assouad's lemma with these two results and choosing $k = n^{1/(2\alpha+1)}$ give immediately that

$$\max_{\Sigma(\theta)\in\mathcal{F}_{11}} 2^2 E_\theta \|\widehat{\Sigma} - \Sigma(\theta)\|^2 \geq \frac{c^2}{2} k^2 a^2 \geq c_1 n^{-2\alpha/(2\alpha+1)}$$

for some universal constant $c_1 > 0$.

3.2 Proof of $\inf_{\widehat{\Sigma}} \sup_{\mathcal{F}_{12}} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|^2 \geq c \log p/n$ using Le Cam's method

For the second lower bound, we will make use of Le Cam's convex hull method.

Lemma 4. *Let X be an observation from a distribution in the collection $\{P_\theta, \theta \in \Theta\}$ where $\Theta = \{\theta_0, \theta_1, \ldots, \theta_{p_1}\}$, and let L be the loss function. Define $r(\theta_0, \theta_m) = \inf_t [L(t, \theta_0) + L(t, \theta_m)]$ and $r_{\min} = \inf_{1\leq m\leq p_1} r(\theta_0, \theta_m)$, and denote $\overline{\mathbb{P}} = \frac{1}{p_1}\sum_{m=1}^{p_1} \mathbb{P}_{\theta_m}$. Let T be an estimator of θ based on an observation from a distribution in the collection $\{P_\theta, \theta \in \Theta = \{\theta_0, \theta_1, \ldots, \theta_{p_1}\}\}$, then*

$$\sup_\theta \mathbb{E}L(T, \theta) \geq \frac{1}{2} r_{\min} \left\|\mathbb{P}_{\theta_0} \wedge \overline{\mathbb{P}}\right\|.$$

For the proof, refer to Section 15.2.2 in Wainwright (2019), or again to the same blog as before.

In order to apply Le Cam's method, we need to first construct a parameter set. For $1 \leq m \leq p_1$, let Σ_m be a diagonal covariance matrix with $\sigma_{mm} = 1 + \sqrt{\tau\frac{\log p_1}{n}}, \sigma_{ii} = 1$ for $i \neq m$, and let Σ_0 be the identity matrix. Let $\mathbf{X}_l = \left(X_1^l, X_2^l, \ldots, X_p^l\right)^T \sim N(0, \Sigma_m)$, and denote the joint density of $\mathbf{X}_1, \ldots, \mathbf{X}_n$ by $f_m, 0 \leq m \leq p_1$ with $p_1 = \max\left\{p, e^{n/2}\right\}$, which can be written as follows:

$$f_m = \prod_{1\leq i\leq n, 1\leq j\leq p, j\neq m} \phi_1\left(x_j^i\right) \cdot \prod_{1\leq i\leq n} \phi_{\sigma_{mm}}\left(x_m^i\right),$$

where $\phi_\sigma, \sigma = 1$ or σ_{mm}, is the density of $N\left(0, \sigma^2\right)$. Denote by f_0 the baseline joint density of $\mathbf{X}_1, \ldots, \mathbf{X}_n$ when $\mathbf{X}_l \sim N(0, \Sigma_0)$.

Let $\theta_m = \Sigma_m$ for $0 \leq m \leq p_1$ and the loss function L be the squared operator norm. It is easy to see $r(\theta_0, \theta_m) = \frac{1}{2}\tau\frac{\log p_1}{n}$ for all $1 \leq m \leq p_1$. Then the lower bound

$$\inf_{\widehat{\Sigma}} \sup_{\mathcal{F}_{12}} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|^2 \geq c \log p/n$$

follows immediately from Le Cam's convex hull lemma if there is a universal constant $c > 0$ such that

$$\left\|\mathbb{P}_{\theta_0} \wedge \overline{\mathbb{P}}\right\| \geq c.$$

Note that for any two densities q_0 and q_1, $\int q_0 \wedge q_1 d\mu = 1 - \frac{1}{2}\int |q_0 - q_1| d\mu$, and Jensen's inequality implies

$$\left[\int |q_0 - q_1| d\mu\right]^2 = \left(\int \left|\frac{q_0 - q_1}{q_1}\right| q_1 d\mu\right)^2 \leq \int \frac{(q_0 - q_1)^2}{q_1} d\mu = \int \frac{q_0^2}{q_1} d\mu - 1.$$

Hence

$$\int q_0 \wedge q_1 d\mu \geq 1 - \frac{1}{2}\left(\int \frac{q_0^2}{q_1} d\mu - 1\right)^{1/2},$$

so that it thus suffices to show that

$$\int \left(\frac{1}{p_1}\sum_{m=1}^{p_1} f_m\right)^2 / f_0 d\mu - 1 \to 0,$$

that is,

$$\frac{1}{p_1^2}\sum_{m=1}^{p_1}\int \frac{f_m^2}{f_0} d\mu + \frac{1}{p_1^2}\sum_{m\neq j}\int \frac{f_m f_j}{f_0} d\mu - 1 \to 0.$$

We now calculate $\int \frac{f_m f_j}{f_0} d\mu$. For $m \neq j$ it is easy to see

$$\int \frac{f_m f_j}{f_0} d\mu - 1 = 0.$$

When $m = j$, we have

$$\begin{aligned}
\int \frac{f_m^2}{f_0} d\mu &= \frac{\left(\sqrt{2\pi\sigma_{mm}}\right)^{-2n}}{(\sqrt{2\pi})^{-n}} \prod_{1\leq i\leq n} \int \exp\left[\left(x_m^i\right)^2\left(-\frac{1}{\sigma_{mm}} + \frac{1}{2}\right)\right] dx_m^i \\
&= \left[1 - (1-\sigma_{mm})^2\right]^{-n/2} = \left(1 - \tau\frac{\log p_1}{n}\right)^{-n/2}.
\end{aligned}$$

Thus

$$\begin{aligned}
&\int \left(\frac{1}{p_1}\sum_{m=1}^{p_1} f_m\right)^2 / f_0 d\mu - 1 \\
&\quad = \frac{1}{p_1^2}\sum_{m=1}^{p_1}\left(\int \frac{f_m^2}{f_0} d\mu - 1\right) \\
&\quad \leq \frac{1}{p_1}\left(1 - \tau\frac{\log p_1}{n}\right)^{-n/2} - \frac{1}{p_1} \\
&\quad = \exp\left[-\log p_1 - \frac{n}{2}\log\left(1 - \tau\frac{\log p_1}{n}\right)\right] - \frac{1}{p_1} \to 0
\end{aligned}$$

for $0 < \tau < 1$, where the last step follows from the inequality $\log(1-x) \geq -2x$ for $0 < x < 1/2$.

References

Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6): 2577 – 2604, 2008. doi: 10.1214/08-AOS600. URL https://doi.org/10.1214/08-AOS600.

T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118 – 2144, 2010. doi: 10.1214/09-AOS752. URL https://doi.org/10.1214/09-AOS752.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer, Dordrecht, 2009. doi: 10.1007/b13794. URL https://cds.cern.ch/record/1315296.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.