Journal of Multivariate Analysis 137 (2015) 161–172

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions

T. Tony Cai [a,*], Tengyuan Liang [a], Harrison H. Zhou [b]

[a] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, United States
[b] Department of Statistics, Yale University, New Haven, CT 06511, United States

ARTICLE INFO

Article history:
Received 30 August 2013
Available online 20 February 2015

AMS 2000 subject classifications:
primary 62H12
62H10
secondary 62F12
94A17

Keywords:
Asymptotic optimality
Central limit theorem
Covariance matrix
Determinant
Differential entropy
Minimax lower bound
Sharp minimaxity

ABSTRACT

Differential entropy and log determinant of the covariance matrix of a multivariate Gaussian distribution have many applications in coding, communications, signal processing and statistical inference. In this paper we consider in the high-dimensional setting optimal estimation of the differential entropy and the log-determinant of the covariance matrix. We first establish a central limit theorem for the log determinant of the sample covariance matrix in the high-dimensional setting where the dimension $p(n)$ can grow with the sample size n. An estimator of the differential entropy and the log determinant is then considered. Optimal rate of convergence is obtained. It is shown that in the case $p(n)/n \to 0$ the estimator is asymptotically sharp minimax. The ultra-high-dimensional setting where $p(n) > n$ is also discussed.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

The determinant of a random matrix is an important functional that has been actively studied in random matrix theory under different settings. See, for example, [15,18,19,12–14,11,27,28,25,23]. In particular, central limit theorems for the log-determinant have been established for random Gaussian matrices in [15], for general real i.i.d. random matrices in [23] under an exponential tail condition on the entries, and for Wigner matrices in [28]. The determinant of random matrices has many applications. For example, the determinant is needed for computing the volume of random parallelotopes, which is of significant interest in random geometry (see [20,24]). More specifically, let $Z = (Z_1, \dots, Z_p)$ be linearly independent random vectors in $\mathbb{R}^n$ with $p \leq n$. Then the convex hull of these p points in $\mathbb{R}^n$ almost surely determines a p-parallelotope and the volume of this random p-parallelotope is given by $\nabla_{n,p} = \det(Z^T Z)^{1/2}$, the squared root of the determinant of th random matrix $Z^T Z$.

The differential entropy and the determinant of the covariance matrix of a multivariate Gaussian distribution play a particularly important role in information theory and statistical inference. The differential entropy has a wide range of applications in many areas including coding, machine learning, signal processing, communications, biosciences and chemistry.

[*] Corresponding author.
E-mail address: tcai@wharton.upenn.edu (T.T. Cai).

http://dx.doi.org/10.1016/j.jmva.2015.02.003
0047-259X/© 2015 Elsevier Inc. All rights reserved.

See [26,16,4,10,21]. For example, in molecular biosciences, the evaluation of entropy of a molecular system is important for understanding its thermodynamic properties. In practice, measurements on macromolecules are often modeled as Gaussian vectors. For a multivariate Gaussian distribution $\mathcal{N}_p(\mu, \Sigma)$, it is well-known that the differential entropy $\mathcal{H}(\cdot)$ is given by

$$\mathcal{H}(\Sigma) = \frac{p}{2} + \frac{p\log(2\pi)}{2} + \frac{\log\det\Sigma}{2}. \tag{1}$$

In this case, estimation of the differential entropy of the system is thus equivalent to estimation of the log determinant of the covariance matrix from the sample. For other applications, the relative entropy (a.k.a. the Kullback–Leibler divergence), which involves the difference of the log determinants of two covariance matrices in the Gaussian case, is important. The determinant of the covariance matrices is also needed for constructing hypothesis tests in multivariate statistics (see [2,22]). For example, the likelihood ratio test for testing linear hypotheses about regression coefficients in MANOVA is based on the ratio of the determinants of two sample covariance matrices [2]. In addition, quadratic discriminant analysis, which is an important technique for classification, requires the knowledge of the difference of the log determinants of the covariance matrices of Gaussian distributions. For these applications, it is important to understand the properties of the log determinant of the sample covariance matrix. The high-dimensional setting where the dimension $p(n)$ grows with the sample size n is of particular current interest.

Motivated by the applications mentioned above, in the present paper we first study the limiting law of the log determinant of the sample covariance matrix for the high-dimensional Gaussian distributions. Let $X_1, \ldots, X_{n+1}$ be an independent random sample from the p-dimensional Gaussian distribution $\mathcal{N}_p(\mu, \Sigma)$. The sample covariance matrix is

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n+1}(X_k - \bar{X})(X_k - \bar{X})^T. \tag{2}$$

A central limit theorem is established for the log determinant of $\hat{\Sigma}$ in the high-dimensional setting where the dimension p grows with the sample size n with the only restriction that $p(n) \le n$. In the case when $\lim_{n\to\infty}\frac{p(n)}{n} = r$ for some $0 \le r < 1$, the central limit theorem shows

$$\frac{\log\det\hat{\Sigma} - \sum_{k=1}^{p}\log\left(1 - \frac{k}{n}\right) - \log\det\Sigma}{\sqrt{-2\log\left(1 - \frac{p}{n}\right)}} \xrightarrow{L} \mathcal{N}(0, 1) \quad \text{as } n \to \infty. \tag{3}$$

The result for the boundary case $p = n$ yields

$$\frac{\log\det\hat{\Sigma} - \log(n-1)! + n\log n - \log\det\Sigma}{\sqrt{2\log n}} \xrightarrow{L} \mathcal{N}(0, 1), \quad \text{as } n \to \infty. \tag{4}$$

In particular, this result recovers the central limit theorem for the log determinant of a random matrix with i.i.d. standard Gaussian entries. See [15,23].

We then consider optimal estimation of the differential entropy and the log-determinant of the covariance matrix in the high-dimensional setting. In the conventional fixed-dimensional case, estimation of the differential entropy has been considered by using both Bayesian and frequentist methods. See, for example, [21,26,1]. A Bayesian estimator was proposed in [26] using the inverse Wishart prior which works without the restriction that dimension is smaller than the sample size. However, how to choose good parameter values for the inverse Wishart prior remains an open question when the population covariance matrix is nondiagonal. A uniformly minimum variance unbiased estimator (UMVUE) was constructed in [1]. It was later proved in [21] that this UMVUE is in fact dominated by a Stein-type estimator and is thus inadmissible. The construction of an admissible estimator was left as an open problem in [21].

Based on the central limit theorem for the log determinant of the sample covariance matrix $\hat{\Sigma}$, we consider an estimator of the differential entropy and the log determinant of Σ and study its properties. A non-asymptotic upper bound for the mean squared error of the estimator is obtained. To show the optimality of the estimator, non-asymptotic minimax lower bounds are established using Cramer–Rao's information inequality. The lower bound results show that consistent estimation of $\log\det\Sigma$ is only possible when $\frac{p(n)}{n} \to 0$. Furthermore, it is shown that the estimator is asymptotically sharp minimax in the setting of $\frac{p(n)}{n} \to 0$.

The ultra-high-dimensional setting where $p(n) > n$ is important due to many contemporary applications. It is a common practice in high-dimensional statistical inference, including compressed sensing and covariance matrix estimation, to impose structural assumption such as sparsity on the target in order to effectively estimate the quantity of interest. It is of significant interest to consider estimation of the log determinant of the covariance matrix and the differential entropy in the case $p(n) > n$ under such structural assumptions. A minimax lower bound is given in Section 4 using Le Cam's method which shows that it is in fact not possible to estimate the log determinant consistently even when the covariance matrix is known to be diagonal with equal values. This negative result implies that consistent estimation of $\log\det\Sigma$ is not possible when $p(n) > n$ over all the collections of the commonly considered structured covariance matrices such as bandable, sparse, or Toeplitz covariance matrices.

The rest of the paper is organized as follows. Section 2 establishes a central limit theorem for the log determinant of the sample covariance matrix. Section 3 considers optimal estimation of the differential entropy and the log-determinant of the covariance matrix. The optimal rate of convergence is established and the estimator is shown to be asymptotically sharp minimax when $\frac{p(n)}{n} \to 0$. Section 4 discusses related applications and the case of $p(n) > n$. The proofs of the main results are given in Section 5.

2. Limiting law of the log determinant of the sample covariance matrix

In this section, we consider the limiting distribution of the log determinant of the sample covariance matrix $\hat{\Sigma}$ and establish a central limit theorem for $\log\det\hat{\Sigma}$ in the high-dimensional setting where $p(n)$ can grow with n under the restriction that $p(n) \le n$.

For two positive integers n and p, define the constant $\tau_{n,p}$ by

$$\tau_{n,p} := \sum_{k=1}^{p}\left[\psi\left(\frac{n-k+1}{2}\right) - \log\left(\frac{n}{2}\right)\right] \tag{5}$$

where $\psi(x) = \frac{\partial}{\partial z}\log\Gamma(z)\big|_{z=x}$ is the Digamma function with $\Gamma(z)$ being the gamma function, and define the constant $\sigma_{n,p}$ by

$$\sigma_{n,p} := \left(\sum_{k=1}^{p}\frac{2}{n-k+1}\right)^{\frac{1}{2}}. \tag{6}$$

We have the following central limit theorem for $\log\det\hat{\Sigma}$.

Theorem 1 (*Asymptotic Distribution*). *Let $X_1,\ldots,X_{n+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu,\Sigma)$. Suppose that $n \to \infty$ and $p(n) \le n$. Then the log determinant of the sample covariance matrix $\hat{\Sigma}$ satisfies*

$$\frac{\log\det\hat{\Sigma} - \tau_{n,p} - \log\det\Sigma}{\sigma_{n,p}} \xrightarrow{L} \mathcal{N}(0,1) \quad \text{as } n\to\infty, \tag{7}$$

where the constants $\tau_{n,p}$ and $\sigma_{n,p}$ are given in (5) and (6) respectively.

Note that Theorem 1 holds with either p fixed or $p(n)$ growing with n, as long as $p(n) \le n$. The assumption in Theorem 1 is generally mild. For example, it does not require that the limit of the ratio $\frac{p(n)}{n}$ exists. In particular, the theorem covers the following four special settings: (1) Fixed p; (2) $\lim_{n\to\infty}\frac{p(n)}{n} = r$ for some $0 \le r < 1$; (3) $p(n) < n$ and $\lim_{n\to\infty}\frac{p(n)}{n} = 1$; (4) The boundary case $p(n) = n$.

It is helpful to look at these special cases separately. Case (1) with fixed p is the classical setting. In this case, asymptotic normality of the determinant $\det\hat{\Sigma}$ has been well studied [2,22]. For completeness, we state the result for $\log\det\hat{\Sigma}$ below.

Corollary 1 (*Case (1): Fixed p*). *If p is fixed, then the log determinant of $\hat{\Sigma}$ satisfies*

$$\frac{\log\det\hat{\Sigma} - p(p+1)/(2n) - \log\det\Sigma}{\sqrt{2p/n}} \xrightarrow{L} \mathcal{N}(0,1), \quad \text{as } n\to\infty. \tag{8}$$

We now consider Case (2) where $\lim_{n\to\infty}\frac{p(n)}{n} = r$ for some $0 \le r < 1$. It is easy to verify that in this case the constants $\tau_{n,p}$ and $\sigma_{n,p}$ satisfy

$$\tau_{n,p} = \sum_{k=1}^{p}\log\left(1-\frac{k}{n}\right) + O\left(\frac{1}{n}\right) \quad\text{and}\quad \sigma_{n,p} = \sqrt{-2\log\left(1-\frac{p}{n}\right)} + O\left(\frac{1}{n}\right). \tag{9}$$

It can be seen easily that $\tau_{n,p} \to -\infty$ at the rate $O(n)$ when $0 < r < 1$. We have the following corollary for Case (2), which reduces to [17].

Corollary 2 (*Case (2): $0 \le r < 1$*). *If $\lim_{n\to\infty}\frac{p(n)}{n} = r$ for some $0 \le r < 1$, then the log determinant $\log\det\hat{\Sigma}$ satisfies*

$$\frac{\log\det\hat{\Sigma} - \sum_{k=1}^{p}\log(1-k/n) - \log\det\Sigma}{\sqrt{-2\log(1-p/n)}} \xrightarrow{L} \mathcal{N}(0,1) \quad \text{as } n\to\infty. \tag{10}$$

Case (3) is more complicated. Unlike the other three cases, it cannot be reduced to a simpler form than the original Theorem 1 in general. We consider two interesting special settings: (a) $\frac{p(n)}{n} \to 1$ and $n - p(n) \to \infty$; (b) $n - p(n)$ is uniformly bounded.

In case (a), the central limiting theorem is of the same form as in Corollary 2. In case (b), the central limiting theorem is of the same form as the boundary case of $p(n) = n$ which is given as follows.

Corollary 3 (*Boundary Case:* $p(n) = n$)**.** *If* $p(n) = n$, *the log determinant* $\log\det\hat{\Sigma}$ *satisfies*

$$\frac{\log\det\hat{\Sigma} - \log(n-1)! + n\log n - \log\det\Sigma}{\sqrt{2\log n}} \xrightarrow{L} \mathcal{N}(0,1), \quad \text{as } n \to \infty. \tag{11}$$

It is interesting to note that the result given in (11) for the boundary case $p(n) = n$ in fact recovers the central limit theorem for the log determinant of a random Gaussian matrix $Y = (y_{ij})_{n\times n}$ with i.i.d. $\mathcal{N}(0,1)$ entries y_{ij},

$$\frac{\log|\det Y| - \frac{1}{2}\log(n-1)!}{\sqrt{\frac{1}{2}\log n}} \xrightarrow{L} \mathcal{N}(0,1), \quad \text{as } n \to \infty. \tag{12}$$

See, for example, [15,23]. This can be seen as follows. When $p = n$, it can be verified directly that the log determinant $\log\det\hat{\Sigma}$ satisfies

$$\log\det\hat{\Sigma} + n\log n - \log\det\Sigma = \log\det(Y^TY) = 2\log|\det Y| \tag{13}$$

where Y is an $n \times n$ random matrix whose entries are independent standard Gaussian variables. Thus Corollary 3 yields

$$\frac{2\log|\det Y| - \log(n-1)!}{\sqrt{2\log n}} \xrightarrow{L} \mathcal{N}(0,1), \tag{14}$$

which is equivalent to (12).

3. Estimation of log-determinant and differential entropy

As mentioned in Introduction, the log-determinant of the covariance matrix and differential entropy is important in many applications. In this section, we consider optimal estimation of the log-determinant and differential entropy of high-dimensional Gaussian distributions. Both minimax upper and lower bounds are given and the results yield sharp asymptotic minimaxity.

Suppose we observe $X_1, \ldots, X_{n+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$. Based on the central limit theorem for $\log\det\hat{\Sigma}$ given in Theorem 1, we consider the following estimator for the log determinant $T = \log\det\Sigma$ of the covariance matrix Σ,

$$\hat{T} = \log\det\hat{\Sigma} - \tau_{n,p} \tag{15}$$

and the corresponding estimator of the differential entropy $\mathcal{H}(\Sigma)$ given by

$$\widehat{\mathcal{H}(\Sigma)} = \frac{p}{2} + \frac{p\log(2\pi)}{2} + \frac{\log\det\hat{\Sigma}}{2} - \frac{\tau_{n,p}}{2}. \tag{16}$$

Here the constant $\tau_{n,p}$ as defined in (5) can be viewed as a bias correction term. The estimators (15) and (16) have been studied in [1,21] in the fixed-dimensional setting. It was shown to be a UMVUE in [1] and inadmissible in [21]. When the dimension p is fixed, the bias correction term $\tau_{n,p}$ is of order $\frac{1}{n}$ and is thus negligible. In particular, the log-determinant of the sample covariance matrix $\log\det\hat{\Sigma}$ is asymptotically unbiased as an estimator of $\log\det\Sigma$. Here we consider the estimator in the high-dimensional setting where the dimension p can grow with n under the only restriction $p(n) \le n$. The bias correction term $\tau_{n,p}$ plays a much more prominent role in such a setting because as discussed in Section 2, $\tau_{n,p}$ is of order n when $\lim_{n\to\infty}\frac{p(n)}{n} = r$ for some $0 < r < 1$.

In this section, we focus on the asymptotic behavior and optimality of the estimators $\hat{T}$ and $\widehat{\mathcal{H}(\Sigma)}$. We establish a non-asymptotic upper bound for mean square error, a minimax lower bound and the optimal rate of convergence as well as sharp asymptotic minimaxity for the estimators $\hat{T}$ and $\widehat{\mathcal{H}(\Sigma)}$ in the following two subsections. Since the log determinant $\log\det\Sigma$ and the differential entropy $\mathcal{H}(\Sigma)$ only differ by a constant in the Gaussian case, the two estimation problems are essentially the same. We shall focus on estimation of $\log\det\Sigma$ in the rest of this section.

3.1. Upper bound

We begin by giving a non-asymptotic upper bound for the mean squared error of the estimator $\hat{T}$.

Theorem 2 (*Non-Asymptotic Upper Bound*)**.** *Suppose* $p \le n$. *Let the estimator* $\hat{T}$ *be defined in* (15). *Then the risk of* $\hat{T}$ *satisfies*

$$\mathbb{E}\left(\hat{T} - \log\det\Sigma\right)^2 \le -2\log\left(1 - \frac{p}{n}\right) + \frac{10p}{3n}\cdot\frac{1}{n-p}. \tag{17}$$

The proof of this theorem is connected to that of Theorem 1 as it can be seen intuitively that

$$\mathbb{E}\left(\hat{T} - \log\det\Sigma\right)^2 \sim \sigma^2_{n,p} = \sum_{k=1}^{p} \frac{2}{n-k+1} \leq -2\log\left(1-\frac{p}{n}\right) \tag{18}$$

which yields the dominate term in (17). The higher order term on the right hand side of (17) can be worked out explicitly using the Taylor expansion with the remainder term. The detailed proof including derivation of the higher order term is given in Section 5.

3.2. Asymptotic optimality

Theorem 2 gives an upper bound for the risk of the estimator $\hat{T}$. We now establish the optimal rate of convergence for estimating $\log\det\Sigma$ by obtaining a minimax lower bound using the information inequality. The results show that the estimator $\hat{T}$ is asymptotically sharp minimax in the case $\lim_{n\to\infty}\frac{p(n)}{n} = 0$.

Theorem 3 (*Non-Asymptotic Information Bound*)**.** *Let* $X_1, \ldots, X_{n+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$. *Suppose* $p \leq n$. *Then the minimax risk for estimating* $\log\det\Sigma$ *satisfies*

$$\inf_{\delta}\sup_{\Sigma}\mathbb{E}(\delta - \log\det\Sigma)^2 \geq 2\cdot\frac{p}{n} \tag{19}$$

where the infimum is taken over all measurable estimators δ *and the supreme is taken over all the possible positive definite covariance matrix* Σ.

The proof of Theorem 3 is given in Section 5. A major tool is the Cramer–Rao information inequality. Together with the upper bound given in (17), we have the following asymptotic optimality result.

Theorem 4 (*Asymptotic Optimality*)**.** *Let* $X_1, \ldots, X_{n+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$. *Suppose that* $n \to \infty$, $p(n) \leq n$ *and* $n - p(n) \to \infty$. *Then*

$$2\cdot\underline{\lim}_{n\to\infty}\frac{p}{n} \leq \underline{\lim}_{n\to\infty}\inf_{\delta}\sup_{\Sigma}\mathbb{E}(\delta - \log\det\Sigma)^2 \leq \overline{\lim}_{n\to\infty}\inf_{\delta}\sup_{\Sigma}\mathbb{E}(\delta - \log\det\Sigma)^2 \leq \overline{\lim}_{n\to\infty}\left(-2\log\left(1-\frac{p}{n}\right)\right). \tag{20}$$

In particular, if $\frac{p(n)}{n} \to 0$, *then the minimax risk satisfies*

$$\lim_{n\to\infty}\frac{n}{p}\cdot\inf_{\delta}\sup_{\Sigma}\mathbb{E}(\delta - \log\det\Sigma)^2 = 2 \tag{21}$$

and the estimator $\hat{T}$ *defined in* (15) *is asymptotically sharp minimax.*

Assume $\lim_{n\to\infty}\frac{p(n)}{n} = r \in [0, 1)$. In the case of $r = 0$, Theorem 4 shows that the optimal constant in the asymptotic risk is 2 and that the estimator $\hat{T}$ given in (15) attains both the optimal rate and the optimal constant asymptotically. It is thus asymptotically sharp minimax. When $0 < r < 1$, the theorem also shows that the minimax risk is non-vanishing and is bounded between $2r$ and $-2\log(1-r)$. It is thus not possible to estimate $\log\det\Sigma$ consistently under the squared error loss in this case.

Remark 1. We have focused on the case $0 \leq r < 1$ in Theorem 4. When $r = 1$, Theorem 3 shows that

$$\underline{\lim}_{n\to\infty}\inf_{\delta}\sup_{\Sigma}\mathbb{E}(\delta - \log\det\Sigma)^2 \geq 2. \tag{22}$$

So consistent estimation of $\log\det\Sigma$ under mean squared error is not possible. If $r = 1$ and $n - p$ is uniformly bounded, then

$$2\cdot\frac{p}{n} \leq \inf_{\delta}\sup_{\Sigma}\mathbb{E}(\delta - \log\det\Sigma)^2 \leq c\log n$$

for some positive constant c, which can be taken as 2 as $n \to \infty$.

In terms of estimating the differential entropy $\mathcal{H}(\Sigma)$, the entropy estimator $\widehat{\mathcal{H}(\Sigma)}$ defined in (16) is asymptotic optimal when $\frac{p(n)}{n} \to 0$, which means that in the asymptotic sense, $\widehat{\mathcal{H}(\Sigma)}$ is the optimal minimax estimator.

4. Discussions

In this paper, we have focused on estimating the log determinant in the "moderately" high-dimensional setting under the restriction that $p(n) \leq n$. The lower bound given in Theorem 3 shows that it is not possible to estimate the log determinant consistently when $\frac{p(n)}{n} \to r > 0$. It is a common practice in high-dimensional statistical inference to impose structural assumption such as sparsity on the parameter in order to effectively estimate the quantity of interest. In the context of covariance matrix estimation, commonly considered collections include bandable covariance matrices, sparse covariance matrices, and Toeplitz covariance matrix. See, for example, [7,9], and [6]. It is interesting to see if the log determinant can be well estimated in the high-dimensional case with $p(n) > n$ under one of these structural constraints. The answer is unfortunately negative.

For any constant $K > 1$, define the following collection of p-dimensional bounded diagonal covariance matrices,

$$\mathcal{D}_K = \{\mathrm{diag}(\overbrace{a, a, \ldots, a}^{p}) : 1/K \leq a \leq K\}. \tag{23}$$

When $p(n) > n$, the following minimax lower bound shows that it is not possible to accurately estimate the log determinant even for the simple diagonal matrices in $\mathcal{D}_K$.

Theorem 5 (*Minimax Lower Bounds*). *Let $X_1, \ldots, X_{n+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$. The minimax risk of estimating the log determinant of the covariance matrix Σ over the collection $\mathcal{D}_K$ of bounded diagonal matrices satisfies,*

$$\inf_{\delta} \sup_{\Sigma \in \mathcal{D}_K} \mathbb{E}\left(\delta - \log\det\Sigma\right)^2 \geq C_K \cdot \frac{p}{n}, \tag{24}$$

for all n, p, where C_K is a constant that satisfies $0 < C_K \leq 2$.

The proof of this minimax lower bound is given in Section 5 using Le Cam's method. Theorem 5 shows that when $p(n) > n$ it is not possible to estimate consistently the bounded diagonal matrices in $\mathcal{D}_K$. Since all the reasonable collections of covariance matrices including the three collections mentioned earlier contain $\mathcal{D}_K$ as a subset, it is thus also impossible to estimate $\log\det\Sigma$ consistently over those commonly used collection of covariance matrices when the dimension is larger than the sample size.

In addition to the differential entropy considered in this paper, estimating the log determinant of covariance matrices is needed for many other applications. One common problem in statistics and engineering is to estimate the distance between two population distributions based on the samples. A commonly used measure of closeness is the relative entropy or the Kullback–Leibler divergence. For two distributions $\mathbb{P}$ and $\mathbb{Q}$ with respective density functions $p(\cdot)$ and $q(\cdot)$, the relative entropy between $\mathbb{P}$ and $\mathbb{Q}$ is

$$KL(\mathbb{P}, \mathbb{Q}) = \int p(x) \log \frac{p(x)}{q(x)} dx. \tag{25}$$

In the case of two multivariate Gaussian distributions $\mathbb{P} = \mathcal{N}_p(\mu_1, \Sigma_1)$ and $\mathbb{Q} = \mathcal{N}_p(\mu_2, \Sigma_2)$,

$$KL(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_2^{-1}\Sigma_1\right) - p + (\mu_2 - \mu_1)^T \Sigma_2^{-1}(\mu_2 - \mu_1) + \log\left(\frac{\det\Sigma_1}{\det\Sigma_2}\right)\right). \tag{26}$$

From (26), it is clear that estimation of the relative entropy involves estimation of the log determinants $\log\det\Sigma_1$ and $\log\det\Sigma_2$. The results given in this paper can be readily used for this part of the estimation problem.

The estimation results obtained in the present paper can also be applied for testing the hypothesis that two multivariate Gaussian distributions $\mathbb{P} = \mathcal{N}_p(\mu_1, \Sigma_1)$ and $\mathbb{Q} = \mathcal{N}_p(\mu_2, \Sigma_2)$ have the same entropy,

$$H_0 : \mathcal{H}(\mathbb{P}) = \mathcal{H}(\mathbb{Q}) \quad \text{vs.}\ H_1 : \mathcal{H}(\mathbb{P}) \neq \mathcal{H}(\mathbb{Q}). \tag{27}$$

For any given significance level $0 < \alpha < 1$, a test with the asymptotic level α can be easily constructed using the central limit theorem given in Section 2, based on two independent samples, one from $\mathbb{P}$ and another from $\mathbb{Q}$.

Knowledge of the log determinant of covariance matrices is also essential for the quadratic discriminant analysis (QDA). For classification of two multivariate Gaussian distributions $\mathcal{N}_p(\mu_1, \Sigma_1)$ and $\mathcal{N}_p(\mu_2, \Sigma_2)$, when the parameters μ_1, μ_2, Σ_1 and Σ_2 are known, the oracle discriminant is

$$\Delta = -(z - \mu_1)^T \Sigma_1^{-1}(z - \mu_1) + (z - \mu_2)^T \Sigma_2^{-1}(z - \mu_2) - \log\left(\frac{\det\Sigma_1}{\det\Sigma_2}\right). \tag{28}$$

That is, the observation z is classified into the population with $\mathcal{N}_p(\mu_1, \Sigma_1)$ distribution if $\Delta > 0$ and into $\mathcal{N}_p(\mu_2, \Sigma_2)$ otherwise. In applications, the parameters are unknown and the oracle discriminant needs to be estimated from data. One of the important quantities in (28) involves the log determinants. Efficient estimation of $\log\det\Sigma_1 - \log\det\Sigma_2$ leads to a better QDA rule.

5. Proofs

We give the proofs of the main results in this section. We begin by collecting two basic but important lemmas for the proof of Theorem 2.

Lemma 1. *Let $X_1, \ldots, X_{n+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$ with $p \le n$. Denote the sample covariance matrix by $\hat{\Sigma}$. Then*

$$\log\det\hat{\Sigma} - \log\det\Sigma = \log\det\hat{I} \tag{29}$$

where $\hat{I} = \frac{1}{n}\sum_{k=1}^{n} Y_k Y_k^T$ is the sample covariance matrix for independent and identically distributed p-variate Gaussian random variables $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0, I)$, I is the identity matrix.

Proof. Note that the distribution of $\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n+1}(X_k - \bar{X})(X_k - \bar{X})^T$ is the same as $\frac{1}{n}\sum_{k=1}^{n} Z_k Z_k^T$, where $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0, \Sigma)$. See, for example, [2]. Define $Y_k = \Sigma^{-1/2} Z_k$, then $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0, I)$ and

$$\begin{aligned}
\log\det\hat{\Sigma} - \log\det\Sigma &= \log\det\left(\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}\right) \\
&= \log\det\left(\Sigma^{-1/2}\left(\frac{1}{n}\sum_{k=1}^{n} Z_k Z_k^T\right)\Sigma^{-1/2}\right) \\
&= \log\det\left(\frac{1}{n}\sum_{k=1}^{n} Y_k Y_k^T\right) \\
&= \log\det\hat{I}. \quad \blacksquare
\end{aligned}$$

A variant of the well-known Bartlett decomposition [2] in multivariate statistics implies the following lemma on the distribution of the determinant of the sample covariance matrix.

Lemma 2. *The law of $\log\det\left(n\hat{I}\right)$ is the same as the sums of p-independent $\log\chi^2$ distribution, namely*

$$\log\det\left(n\hat{I}\right) \overset{L}{=} \sum_{k=1}^{p}\log\left(\chi^2_{n-k+1}\right) \tag{30}$$

where $\chi^2_n, \ldots, \chi^2_{n-p+1}$ are mutually independent χ^2 distribution with the degrees of freedom $n, \ldots, n-p+1$ respectively.

5.1. *Proof of Theorem 1*

The proof of Theorem 1 relies on the above two lemmas, the following Lemma 3 and an analysis of the characteristic functions.

Lemma 3.

$$r_{n,p} = \frac{\sum_{k=1}^{p}\frac{1}{(n-k+1)^2}}{\sum_{k=1}^{p}\frac{1}{n-k+1}} \le \max\left\{\frac{1}{\log n + 1}, \frac{\frac{\pi^2}{6}}{\log(n+1) - \log(\log n + 1)}\right\} \to 0 \tag{31}$$

uniformly in $p(n)$ as $n \to \infty$, where $p(n)$ can grow with n, $p(n) \le n$.

Proof. We consider the following two scenarios (1) when $n - p(n) \ge \log n$. (2) when $n - p(n) \le \log n$.

For case (1), Eq. (31) can be bounded in the following way

$$\begin{aligned}
r_{n,p} &= \frac{\sum_{k=1}^{p}\frac{1}{(n-k+1)^2}}{\sum_{k=1}^{p}\frac{1}{n-k+1}} \\
&\le \frac{\frac{1}{n-p+1}\cdot\sum_{k=1}^{p}\frac{1}{n-k+1}}{\sum_{k=1}^{p}\frac{1}{n-k+1}} \le \frac{1}{\log n + 1}.
\end{aligned} \tag{32}$$

For case (2), Eq. (31) can be bounded in the following way

$$
\begin{aligned}
r_{n,p} &= \frac{\sum_{k=1}^{p} \frac{1}{(n-k+1)^2}}{\sum_{k=1}^{p} \frac{1}{n-k+1}} \\
&\le \frac{\frac{\pi^2}{6}}{\sum_{k=1}^{p} \log\left(1+\frac{1}{n-k+1}\right)} \\
&\le \frac{\frac{\pi^2}{6}}{\log(n+1)-\log(n-p+1)} \le \frac{\frac{\pi^2}{6}}{\log(n+1)-\log(\log n+1)}.
\end{aligned} \tag{33}
$$

Thus, we have the following bound for $r_{n,p}$ uniformly in $p(n)$

$$
r_{n,p} \le \max\left\{\frac{1}{\log n+1}, \frac{\frac{\pi^2}{6}}{\log(n+1)-\log(\log n+1)}\right\} \to 0, \quad \text{as } n\to\infty. \tag{34}
$$

Basically we show this sequence converges to 0 uniformly faster than the $O(1/\log n)$ rate. ■

It follows from Lemmas 1 and 2 that

$$
\hat{T} - \log\det\Sigma = \log\det\left(n\hat{I}\right) - \sum_{k=1}^{p}\left[\psi\left(\frac{1}{2}(n-k+1)\right)+\log 2\right] \triangleq Z. \tag{35}
$$

Thus,

$$
Z \overset{L}{=} \sum_{k=1}^{p}\left[\log\left(\chi^2_{n-k+1}\right) - \psi\left(\frac{1}{2}(n-k+1)\right) - \log 2\right]. \tag{36}
$$

Inspired by [17] (where a special case of our theorem has been proved under much stronger conditions) and using the fact of the independence and the characteristic function of the logarithm Chi-square distribution, the characteristic function of Z is

$$
\begin{aligned}
\phi_Z(t) &= \prod_{k=1}^{p} \phi_{\log\chi^2_{n-k+1}}(t) \cdot \frac{1}{\exp\left(it\cdot\left[\psi\left(\frac{1}{2}(n-k+1)\right)+\log 2\right]\right)} \\
&= \prod_{k=1}^{p} \mathbb{E}e^{it\log\chi^2_{n-k+1}} \cdot \frac{1}{\exp\left(it\cdot\left[\psi\left(\frac{1}{2}(n-k+1)\right)+\log 2\right]\right)} \\
&= \prod_{k=1}^{p} \mathbb{E}(\chi^2_{n-k+1})^{it} \cdot \frac{1}{\exp\left(it\cdot\psi\left(\frac{1}{2}(n-k+1)\right)\right)\cdot 2^{it}} \\
&= \prod_{k=1}^{p} \frac{\Gamma\left(\frac{1}{2}(n-k+1)+it\right)}{\Gamma\left(\frac{1}{2}(n-k+1)\right)} 2^{it} \cdot \frac{1}{\exp\left(it\cdot\psi\left(\frac{1}{2}(n-k+1)\right)\right)\cdot 2^{it}} \\
&= \prod_{k=1}^{p} \frac{\Gamma\left(\frac{1}{2}(n-k+1)+it\right)}{\Gamma\left(\frac{1}{2}(n-k+1)\right)} \cdot \frac{1}{\exp\left(it\cdot\psi\left(\frac{1}{2}(n-k+1)\right)\right)}.
\end{aligned} \tag{37}
$$

Thus we have,

$$
\log\phi_Z(t) = \sum_{k=1}^{p}\left\{\log\Gamma\left(\frac{1}{2}(n-k+1)+it\right) - \log\Gamma\left(\frac{1}{2}(n-k+1)\right) - it\cdot\psi\left(\frac{1}{2}(n-k+1)\right)\right\}.
$$

Using the Taylor expansion of gamma and digamma function [3], we have

$$
\begin{aligned}
\log\Gamma(z) &= z\log z - z - \frac{1}{2}\log\frac{z}{2\pi} + \frac{1}{12z} + O\left(\frac{1}{|z|^2}\right) \\
\psi(z) &= \log z - \frac{1}{2z} + O\left(\frac{1}{|z|^2}\right).
\end{aligned}
$$

Thus for each term in above characteristic function, we have

$$\begin{aligned}
&\log\Gamma\left(\frac{1}{2}(n-k+1)+it\right)-\log\Gamma\left(\frac{1}{2}(n-k+1)\right)-it\cdot\psi\left(\frac{1}{2}(n-k+1)\right)\\
&\quad= it\log\left(\frac{1}{2}(n-k)+1\right)-it\frac{1}{n-k+1}+(it)^2\frac{1}{n-k+1}-it\log\left(\frac{1}{2}(n-k+1)\right)\\
&\qquad+it\frac{1}{n-k+1}+O\left(\frac{1}{(n-k+1)^2}\right)\\
&\quad=(it)^2\frac{1}{n-k+1}+O\left(\frac{|t|^2}{(n-k+1)^2}\right).
\end{aligned} \tag{38}$$

The characteristic function $\phi_0(t)$ of $\frac{1}{\sigma_{n,p}}\left(\log\det\hat{\Sigma}-\tau_{n,p}-\log\det\Sigma\right)$ is

$$\phi_0(t)=\exp\left\{\frac{(it)^2}{2}+O\left(|t|^2\cdot\frac{\sum_{k=1}^{p}\frac{1}{(n-k+1)^2}}{\sum_{k=1}^{p}\frac{1}{n-k+1}}\right)\right\}$$

Lemma 3 shows that

$$r_n=\frac{\sum_{k=1}^{p}\frac{1}{(n-k+1)^2}}{\sum_{k=1}^{p}\frac{1}{n-k+1}}\to 0 \tag{39}$$

under $n\to\infty$ and $p\le n$. Thus, when $n\to\infty$, $\phi_0(t)\to e^{\frac{(it)^2}{2}}$ and the result follows. ■

5.2. Proof of Theorem 2

It follows from the variance of the logarithm Chi-square distribution and the Taylor expansion for TriGamma function that

$$\psi'(z)=\frac{1}{z}+\left(\frac{1}{2z^2}+\frac{1}{6z^3}\right)\theta$$

for $z\ge 1$ and $0<\theta<1$. Hence,

$$\begin{aligned}
\mathbb{E}\left(\hat{T}-\log\det\Sigma\right)^2&=\operatorname{Var}\left(\sum_{k=1}^{p}\log(\chi^2_{n-k+1})\right)\\
&=\sum_{k=1}^{p}\psi'\left(\frac{n-k+1}{2}\right)\\
&=\sum_{k=1}^{p}\left[\frac{2}{n-k+1}+\frac{2\theta}{(n-k+1)^2}+\frac{4\theta}{3(n-k+1)^3}\right]\\
&\le\sum_{k=1}^{p}\left[-2\log\left(1-\frac{1}{n-k+1}\right)+\frac{10}{3(n-k+1)^2}\right].
\end{aligned} \tag{40}$$

Since $\sum_{k=1}^{p}\frac{1}{(n-k+1)^2}\le\sum_{k=1}^{p}\frac{1}{(n-k)(n-k+1)}=\sum_{k=1}^{p}\left(\frac{1}{n-k}-\frac{1}{n-k+1}\right)=\frac{1}{n-p}-\frac{1}{n}$ and $\sum_{k=1}^{p}\log(1-\frac{1}{n-k+1})=\sum_{k=1}^{p}\log(\frac{n-k}{n-k+1})=\log(1-\frac{p}{n})$, we have

$$\begin{aligned}
\mathbb{E}\left(\hat{T}-\log\det\Sigma\right)^2&\le-2\log\left(1-\frac{p}{n}\right)+\frac{10}{3}\cdot\left(\frac{1}{n-p}-\frac{1}{n}\right)\\
&=-2\log\left(1-\frac{p}{n}\right)+\frac{10p}{3n}\cdot\frac{1}{n-p}.\quad■
\end{aligned} \tag{41}$$

5.3. Proof of Theorem 3

We first recall the biased version of Cramer–Rao inequality in the multivariate case. Let $\Theta_{1\times p}$ be a parameter vector and let $X\sim f(\Theta)$, where $f(\Theta)$ is the density function. Consider any estimator $\hat{T}(X)$ of the function $\phi(\Theta)$ with the bias

$B(\Theta) = \mathbb{E}\hat{T}(X) - \phi(\Theta) = (b(\theta_1), \dots, b(\theta_p))^T$. Then

$$\begin{aligned}\mathbb{E}_\Theta(\hat{T} - \phi(\Theta))^2 &= \mathrm{Var}_\Theta(\hat{T}(X)) + \|B(\Theta)\|_2^2 \\ &\geq \left(\frac{\partial\left(\phi(\Theta) + B(\Theta)\right)}{\partial\Theta}\right)^T \cdot [\mathbf{I}(\Theta)]^{-1} \cdot \frac{\partial\left(\phi(\Theta) + B(\Theta)\right)}{\partial\Theta} + B(\Theta)^T \cdot B(\Theta). \qquad (42)\end{aligned}$$

Now consider the diagonal matrix subfamily, $\Sigma = \mathrm{diag}\left(\theta_1, \theta_2, \dots, \theta_p\right)$, with p parameters, $\theta_1, \dots, \theta_p$. We wish to estimate $\phi(\Theta) = \sum_{i=1}^p \log(\theta_i) = \log\det(\Sigma)$. For a random sample $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0, \Sigma)$, the Fisher information matrix and the partial derivative of the $\phi(\Theta)$ are given by

$$\mathbf{I}(\Theta) = \mathrm{diag}\left(\frac{n}{2\theta_1^2}, \frac{n}{2\theta_2^2}, \dots, \frac{n}{2\theta_p^2}\right) \quad \text{and} \quad \frac{\partial\phi(\Theta)}{\partial\Theta} = \left(\frac{1}{\theta_1}, \frac{1}{\theta_2}, \dots, \frac{1}{\theta_p}\right)^T.$$

Eq. (42) can be calculated explicitly as

$$\begin{aligned}&\left(\frac{\partial\left(\phi(\Theta) + B(\Theta)\right)}{\partial\Theta}\right)^T \cdot [\mathbf{I}(\Theta)]^{-1} \cdot \frac{\partial\left(\phi(\Theta) + B(\Theta)\right)}{\partial\Theta} + B(\Theta)^T \cdot B(\Theta) \\ &= \left(\frac{1}{\theta_1} + b'(\theta_1), \dots, \frac{1}{\theta_p} + b'(\theta_p)\right)\left[\mathrm{diag}\left(\frac{n}{2\theta_1^2}, \dots, \frac{n}{2\theta_p^2}\right)\right]^{-1}\left(\frac{1}{\theta_1} + b'(\theta_1), \dots, \frac{1}{\theta_p} + b'(\theta_p)\right)^T + \sum_{k=1}^p b^2(\theta_k) \\ &= \sum_{k=1}^p \left[\frac{2}{n}\left(1 + \theta_k b'(\theta_k)\right)^2 + b^2(\theta_k)\right]. \qquad (43)\end{aligned}$$

As in [5], if we can prove that for any bias function $b(\theta)$

$$\sup_{\theta > 0}\left[\frac{2}{n}\left(1 + \theta b'(\theta)\right)^2 + b^2(\theta)\right] \geq \frac{2}{n}, \qquad (44)$$

then the minimax lower bound result

$$\inf_{\hat{T}} \sup_{\Sigma \in \mathcal{F}} \mathbb{E}(\hat{T} - \log\det\Sigma)^2 \geq 2 \cdot \frac{p}{n} \qquad (45)$$

holds for any parameter space $\mathcal{F}$ containing the set of the diagonal matrices by combining (43) and (44).

To prove Eq. (44), we first prove that for any given constant $K > 0$

$$\sup_{1/K \leq \theta \leq K}\left[\frac{2}{n}\left(1 + \theta b'(\theta)\right)^2 + b^2(\theta)\right] \geq \frac{2}{n}\left(\frac{\log K}{\log K + \sqrt{\frac{2}{n}}}\right)^2. \qquad (46)$$

Assume

$$r_K \geq \sup_{1/K \leq \theta \leq K}\left[\frac{2}{n}\left(1 + \theta b'(\theta)\right)^2 + b^2(\theta)\right],$$

and then we have the following two inequalities

$$|1 + \theta b'(\theta)| \leq \sqrt{\frac{n}{2} \cdot r_K} \quad \text{and} \quad |b(\theta)| \leq \sqrt{r_K},$$

which implies $r_K \geq \frac{2}{n}\left(\frac{\log K}{\log K + \sqrt{\frac{2}{n}}}\right)^2$. This means that $\frac{2}{n}\left(\frac{\log K}{\log K + \sqrt{\frac{2}{n}}}\right)^2$ is a lower bound for (46). Eq. (44) now follows by letting $K \to \infty$. ■

5.4. Proof of Theorem 4

The upper bound given in Theorem 2 yields that

$$\inf_{\delta} \sup_{\Sigma} \mathbb{E}(\delta - \log\det\Sigma)^2 \leq -2 \cdot \log\left(1 - \frac{p}{n}\right) + \frac{p}{n} \cdot \frac{10}{3(n-p)}.$$

It then follows from the assumption $n - p \to \infty$ that

$$\overline{\lim_{n\to\infty}} \inf_{\hat{T}} \sup_{\Sigma} \mathbb{E}(\hat{T} - \log\det\Sigma)^2 \leq 2 \cdot \overline{\lim}_{n\to\infty} - \log\left(1 - \frac{p}{n}\right).$$

When $r = 0$, $-2 \cdot \log\left(1 - \frac{p}{n}\right) \sim 2 \cdot \frac{p}{n}$ and the upper bound follows. For the lower bound, Theorem 3 implies

$$\underline{\lim}_{n\to\infty} \inf_{\hat{T}} \sup_{\Sigma} \mathbb{E}(\hat{T} - \log\det\Sigma)^2 \geq 2 \cdot \underline{\lim}_{n\to\infty} \frac{p}{n}.$$

This completes the proof. ■

5.5. Proof of Theorem 5

The proof uses a two point hypothesis testing argument due to Le Cam (see [29, pp. 79–80]).

Lemma 4 (*Le Cam's Lemma*). *Let $\hat{\theta}$ be any estimator of θ based on an observation from a distribution in the collection $\{\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}\}$, suppose $|\theta_0 - \theta_1| \geq 2s$, and then*

$$\inf_{\hat{\theta}} \sup_{\theta\in\{\theta_0,\theta_1\}} \mathbb{E}(\hat{\theta}_n - \theta)^2 \geq s^2 \cdot \frac{1}{2}\|\mathbb{P}_{\theta_0} \wedge \mathbb{P}_{\theta_1}\| \tag{47}$$

where $\|\mathbb{P} \wedge \mathbb{Q}\| = \int (p \wedge q) d\mu$, is affinity between probability measures.

The total variance affinity can be lower bounded in terms of the χ^2 distance.

Lemma 5 (*Pinsker's Inequality*).

$$\|\mathbb{P} \wedge \mathbb{Q}\| = 1 - TV(\mathbb{P}, \mathbb{Q}) \geq 1 - \sqrt{KL(\mathbb{P}, \mathbb{Q})/2} \geq 1 - \sqrt{\chi^2(\mathbb{P}, \mathbb{Q})/2} \tag{48}$$

where $TV(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\int |p - q| d\mu$ is the total variation distance, $KL(\mathbb{P}, \mathbb{Q})$ is the Kullback–Leibler divergence, $\chi^2(\mathbb{P}, \mathbb{Q})$ is the χ^2 distance.

We use the following lemma, which is a direct consequence of Lemma 2 in [8], to bound the χ^2 distance.

Lemma 6. *For $i = 0$ and 1, let $\mathbb{P}_i$ be the joint distribution of n independent p-dimensional Gaussian variables with the covariance matrix Σ_i. The χ^2 distance $\chi^2(\mathbb{P}_0, \mathbb{P}_1)$ satisfies*

$$\chi^2(\mathbb{P}_0, \mathbb{P}_1) + 1 = \int \frac{P_1^2}{P_0} d\mu = \left\{\det\left(I - (\Sigma_1 - \Sigma_0)\Sigma_0^{-1}(\Sigma_1 - \Sigma_0)\Sigma_0^{-1}\right)\right\}^{-n/2}. \tag{49}$$

To prove the lower bound given in 5, we pick $\Sigma_0 = I_{p\times p}$, $\Sigma_1 = (1 + \frac{1}{\sqrt{np}}) \cdot I_{p\times p}$.

First, let us prove the theorem under $np > \max\{\frac{1}{(K-1)^2}, 1\}$. It is easy to see that these two points lie in the parameter space because $1/K < 1 + \frac{1}{\sqrt{np}} < K$. Then

$$|\theta_0 - \theta_1| = |\log\det\Sigma_0 - \log\det\Sigma_1| = p\log\left(1 + \sqrt{\frac{1}{pn}}\right) > p\frac{\sqrt{\frac{1}{pn}}}{1 + \sqrt{\frac{1}{pn}}} \geq \frac{1}{2}\sqrt{\frac{p}{n}} \tag{50}$$

$$\begin{aligned}\chi^2(\mathbb{P}_0, \mathbb{P}_1) + 1 &= \left\{\det\left(I - (\Sigma_1 - \Sigma_0)\Sigma_0^{-1}(\Sigma_1 - \Sigma_0)\Sigma_0^{-1}\right)\right\}^{-n/2} \\ &= \left(1 - \frac{1}{np}\right)^{-\frac{1}{2}np} < e^{\frac{1}{2}np\frac{\frac{1}{np}}{1-\frac{1}{np}}} < e < \infty \quad \text{For } np > 1.\end{aligned} \tag{51}$$

The χ^2 distance is upper bounded away from infinity; thus the affinity term is lower bounded away from 0. At the same time, the parameters are well separated away with a distance $s = \frac{1}{4}\sqrt{\frac{p}{n}}$. Thus, by Le Cam's lemma, we have, for some constant $c > 0$ ($c \leq 2$ is due to the Theorem 3)

$$\inf_{\delta} \sup_{\Sigma\in\mathcal{D}_K} \mathbb{E}\left(\delta - \log\det\Sigma\right)^2 \geq \left(\frac{1}{4}\sqrt{\frac{p}{n}}\right)^2 \cdot \frac{1}{2}\left(1 - \sqrt{\frac{e-1}{2}}\right) = c \cdot \frac{p}{n}, \tag{52}$$

for all p, n as long as $np > \max\{\frac{1}{(K-1)^2}, 1\}$. More specifically, c can be taken as $\frac{1}{32}\left(1 - \sqrt{\frac{e-1}{2}}\right)$.

Second, for $np \leq \max\{\frac{1}{(K-1)^2}, 1\}$, there is only finite collection of (n, p) pairs; thus we must have a constant c_K small enough such that

$$\inf_{\delta} \sup_{\Sigma\in\mathcal{D}_K} \mathbb{E}\left(\delta - \log\det\Sigma\right)^2 \geq c_K \cdot \frac{p}{n}. \tag{53}$$

Thus combining two parts, we can pick $C_K = \min\{c_K, c\}$, which completes the proof. ■

Acknowledgments

The research of Tony Cai was supported in part by NSF Grant DMS-0604954 and NSF FRG Grant DMS-0854973. The research of Harrison Zhou was supported in part by NSF Career Award DMS-0645676 and NSF FRG Grant DMS-0854975.

References

[1] N.A. Ahmed, D. Gokhale, Entropy expressions and their estimators for multivariate distributions, IEEE Trans. Inform. Theory 35 (3) (1989) 688–692.

[2] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, in: Wiley Series in Probability and Statistics, 2003.

[3] H. Bateman, A. Erdélyi, Higher Transcendental Functions, McGraw-Hill, 1981.

[4] J. Beirlant, E. Dudewicz, L. Györfi, E.C. Van der Meulen, Nonparametric entropy estimation: an overview, Int. J. Math. Stat. Sci. 6 (1997) 17–40.

[5] L.D. Brown, M.G. Low, Information inequality bounds on the minimax risk (with an application to nonparametric regression), Ann. Statist. 19 (1) (1991) 329–337.

[6] T.T. Cai, Z. Ren, H.H. Zhou, Optimal rates of convergence for estimating Toeplitz covariance matrices, Probab. Theory Related Fields (2012) 1–43.

[7] T.T. Cai, C.-H. Zhang, H.H. Zhou, Optimal rates of convergence for covariance matrix estimation, Ann. Statist. 38 (4) (2010) 2118–2144.

[8] T.T. Cai, H.H. Zhou, Minimax estimation of large covariance matrices under $\ell - 1$ norm, Statist. Sinica (2011) (with discussion).

[9] T.T. Cai, H.H. Zhou, Optimal rates of convergence for sparse covariance matrix estimation, Ann. Statist. 40 (5) (2012) 2389–2420.

[10] J.A. Costa, A.O. Hero, Geodesic entropic graphs for dimension and entropy estimation in manifold learning, IEEE Trans. Signal Process. 52 (8) (2004) 2210–2221.

[11] R. Delannay, G. Le Caër, Distribution of the determinant of a random real-symmetric matrix from the Gaussian orthogonal ensemble, Phys. Rev. E 62 (2 Pt A) (2000) 1526.

[12] V.L. Girko, The central limit theorem for random determinants, Theory Probab. Appl. 24 (4) (1980) 729–740.

[13] V.L. Girko, Theory of Random Determinants, Kluwer, Dordrecht, 1990.

[14] V.L. Girko, A refinement of the central limit theorem for random determinants, Theory Probab. Appl. 42 (1) (1998) 121–129.

[15] N. Goodman, The distribution of the determinant of a complex Wishart distributed matrix, Ann. Math. Statist. (1963) 178–180.

[16] M. Gupta, S. Srivastava, Parametric Bayesian estimation of differential entropy and relative entropy, Entropy 12 (4) (2010) 818–843.

[17] D. Jonsson, Some limit theorems for the eigenvalues of a sample covariance matrix, J. Multivariate Anal. 12 (1) (1982) 1–38.

[18] J. Komlós, On the determinant of /0, 1/ matrices, Studia Sci. Math. Hungar. 2 (1) (1967) 7–21.

[19] J. Komlós, On the determinant of random matrices, Studia Sci. Math. Hungar. 3 (1968) 387–399.

[20] A. Mathai, Random p-content of a p-parallelotope in euclidean n-space, Adv. Appl. Probab. 31 (2) (1999) 343–354.

[21] N. Misra, H. Singh, E. Demchuk, Estimation of the entropy of a multivariate normal distribution, J. Multivariate Anal. 92 (2) (2005) 324–342.

[22] R.J. Muirhead, Aspects of Multivariate Statistical Theory, Wiley, 1982.

[23] H.H. Nguyen, V. Vu, Random matrices: law of the determinant, 2011. ArXiv Preprint arXiv:1112.0752.

[24] J. Nielsen, The distribution of volume reductions induced by isotropic random projections, Adv. Appl. Probab. 31 (4) (1999) 985–994.

[25] A. Rouault, Asymptotic behavior of random determinants in the Laguerre, Gram and Jacobi ensembles, 2006. ArXiv Preprint arXiv:math/0607767.

[26] S. Srivastava, M.R. Gupta, Bayesian estimation of the entropy of the multivariate Gaussian, in: IEEE International Symposium on Information Theory, 2008, pp. 1103–1107.

[27] T. Tao, V. Vu, On random ± 1 matrices: singularity and determinant, Random Struct. Algorithms 28 (1) (2006) 1–23.

[28] T. Tao, V. Vu, A central limit theorem for the determinant of a Wigner matrix, Adv. Math. 231 (1) (2012) 74–101.

[29] A.B. Tsybakov, Introduction to Nonparametric Estimation, Springer, 2008.