Bayesian Statistics

31/01/2019

Bernstein-von Mises theorems for functionals of the covariance matrix

Authors:
Hamza Filali Baba
Alice Guichenez
Cédric Vincent-Cuaz
Petr Zamolodtchikov

Professors:
Rémi Bardenet
Anna Simoni

Contents

Introduction

In this paper, we summarize and discuss the article *Bernstein-von Mises theorems for functionals of the covariance matrix* by Chao Gao and Harrison H. Zhou in 2016 [1]. This article extends the Bernstein-von Mises (BvM) theorems for matrix functionals, including matrix entries, quadratic forms, log-determinant and eigenvalues when the dimension of the matrix p grows with the sample size n.

The BvM theorem states that, given a parametric model $(P_\theta, \theta \in \Theta)$, a prior distribution $\theta \sim \Pi$, i.i.d. observations $(X_1, \dots, X_n)$ from the measure $P_{\theta^\star}^n$ and further weak assumptions, the conditional distribution of $\sqrt{n}(\theta - \hat{\theta})|X$ is asymptotically $\mathcal{N}(0, V^2)$ with some centering $\hat{\theta}$ and covariance V^2. Consequently, the distributions $\sqrt{n}(\theta - \hat{\theta})|X^n$ and $\sqrt{n}(\hat{\theta} - \theta)|\theta = \theta^\star$ are asymptotically the same under the sampling distribution $P_{\theta^\star}^n$, which makes the connection between Bayesian and frequentist considerations.

The article considers a Gaussian distribution and puts a prior on the covariance matrix. The conclusions for the matrix functionals are drawn from a general theoretical framework, which they extent to obtain results for linear and quadratic discriminant analysis.

We first emphasize the main points of the paper by making links with the classes and present the main steps of the proofs, then we present the applications of the theorem to specific matrix functionals suggested by the article and propose a theoretical extension, and finally suggest a practical implementation.

1 Summary of the paper and links with classes

The classical BvM result concerns parametrical models, where the parameters are supposed to be finite-dimensional. It shows that the frequentist approach of estimating the parameter (i.e. the MLE) has the same asymptotic distribution as the posterior in the Bayesian fashion.

1.1 Presentation of the results

We first present the results obtained by the authors for Gaussian samples and their illustration to two priors, and then the results they obtain by extending the framework to mixtures of Gaussian samples - for discriminant analysis.

1.1.1 Results within a general scheme

In this section, we consider Gaussian samples, that is i.i.d. samples $X^n = (X_1, \dots, X_n)$ drawn from $\mathcal{N}(0, \Sigma^*)$, where Σ^* is a $p \times p$ matrix with inverse Ω^*. A Bayes method puts a prior Π on the precision matrix Ω, and the posterior distribution is defined as

$$\Pi(B|X^n) = \frac{\int_B \exp(l_n(\Omega)) d\Pi(\Omega)}{\int \exp(l_n(\Omega)) d\Pi(\Omega)},$$

where $l_n(\Omega)$ is the log-likelihood of $\mathcal{N}(0, \Omega^{-1})$ defined as $l_n(\Omega) = \frac{n}{2}\text{logdet}(\Omega) - \frac{n}{2}\text{tr}(\Omega\hat{\Sigma})$ plus or minus a normalizing constant, and where $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$. The objective is to show that $\Pi\left(\sqrt{n}V^{-1}(f(\Omega) - \hat{f})\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$, as $(n,p) \longrightarrow \infty$. In the article, the authors choose to focus on the centering $\hat{f}$ to be the sample version of $f(\Omega) = f(\Sigma^{-1})$, where Σ is replaced by the sample covariance $\hat{\Sigma}$ and they compare the BvM results with the classical asymptotical normality for $\hat{f}$ in the frequentist sense.

Results for approximately linear functionals

Within this framework, the authors present a theorem for approximately linear functionals of the covariance matrix Σ.

We assume there is a set A_n satisfying $A_n \subset \{||\Sigma - \Sigma^*|| \leq \delta_n\}$ for any sequence $\delta_n = o(1)$. We consider a functional of Σ, $f = \phi(\Sigma)$, that is linear in a neighborhood of the truth in the sense that there exists a symmetric matrix Φ such that

$$\sup_{A_n} \sqrt{n}||\Sigma^{*1/2}\Phi\Sigma^{*1/2}||_F^{-1}|\phi(\Sigma) - \phi(\hat{\Sigma}) - \mathrm{tr}((\Sigma - \hat{\Sigma})\Phi)| = o_P(1).$$

Theorem 2.1 *Under the assumptions above and* $\max(||\Sigma^*||, ||\Omega^*||) = O(1)$, *if for a given prior* Π, *the following two conditions are satisfied:*

1. $\Pi(A_n|X^n) = 1 - o_P(1)$,
2. *For any fixed* $t \in \mathbb{R}$, $\frac{\int_{A_n} \exp(l_n(\Omega_t))d\Pi(\Omega)}{\int_{A_n} \exp(l_n(\Omega))d\Pi(\Omega)} = 1 + o_P(1)$ *for the perturbed precision matrix*

$$\Omega_t = \Omega + \frac{\sqrt{2}t}{\sqrt{n}||\Sigma^{*1/2}\Phi\Sigma^{*1/2}||_F}\Phi$$

then

$$\sup_{t\in\mathbb{R}}\left|\Pi\left(\frac{\sqrt{n}(\phi(\Sigma) - \phi(\hat{\Sigma}))}{\sqrt{2}||\Sigma^{*1/2}\Phi\Sigma^{*1/2}||_F} \leq t|X^n\right) - P(Z \leq t)\right| = o_P(1)$$

where $Z \sim \mathcal{N}(0, 1)$.

In other words, that theorem states that under the conditions that (1) the posterior distribution concentrates on a neighbourhood of the truth under the spectral norm on which the functional is approximately linear and (2) the bias caused by the shifted parameter can be absorbed by the posterior distribution, then the asymptotic posterior distribution of $\phi(\Sigma)$ is

$$\mathcal{N}\left(\phi(\hat{\Sigma}), 2n^{-1}||\Sigma^{*1/2}\Phi\Sigma^{*1/2}||_F^2\right).$$

They also present a similar theorem for the precision matrix Ω, which states that under analogous conditions - see section 1.2 for the exact conditions - on linear approximation and on posterior and perturbed precision matrix the asymptotic distribution of $\psi(\Omega)$ is

$$\mathcal{N}\left(\psi(\hat{\Sigma}^{-1}), 2n^{-1}||\Omega^{*1/2}\Psi\Omega^{*1/2}||_F^2\right).$$

Priors

The authors illustrate the above theorem by providing examples of priors and specifying under which conditions that theorem can be applied. Even though the result of a conjugate prior could be derived by direct exploration of the posterior form, the general framework previously defined makes it possible to work with both conjugate and non-conjugate priors. A covariance matrix prior can be considered as a vector prior with an additional constraint of positive semi-definiteness.

Wishart prior

We consider the Wishart prior $\mathcal{W}_p(I, p + b - 1)$ on Ω - symmetric positive semi-definite matrix - with density function

$$\frac{d\Pi(\Omega)}{d\Omega} \quad \alpha \quad \exp\left(\frac{b-2}{2}\log\det(\Omega) - \frac{1}{2}\mathrm{tr}(\Omega)\right).$$

Lemma 2.1 *Assume* $\max(||\Sigma^*||, ||\Omega^*||) = O(1)$ *and* $p/n = o(1)$. *Then, for any integer* $b = O(1)$, *the prior* $\Pi = \mathcal{W}_p(I, p+b-1)$ *satisfies the two conditions in Theorem 1 for some* A_n. *If the extra assumption* $\frac{rp^2}{n} = o(1)$ *is made, the two conditions for the theorem for the precision matrix to hold are also satisfied for some appropriate* A_n.

Gaussian prior

We consider Gaussian prior on Ω with density function

$$\frac{d\Pi(\Omega)}{d\Omega} \quad \alpha \quad \exp\left(\frac{-1}{2}||\Omega||_F^2\right)$$

where $\Omega \in \left\{\Omega = \Omega^T \succeq 0, ||\Omega|| < 2\Lambda, ||\Sigma|| < 2\Lambda\right\}$ for some $\Lambda > 0$.

Lemma 2.2 *Assume* $\max(||\Sigma^*||, ||\Omega^*||) < \Lambda = O(1)$ *and* $\frac{p^2 \log(n)}{n} = o(1)$. *The Gaussian prior* Π *defined above satisfies the two conditions in Theorem 1 for some appropriate* A_n. *If the extra assumption* $\frac{rp^3 \log(n)}{n} = o(1)$ *is made, the two conditions for the theorem for the precision matrix to hold are also satisfied for some appropriate* A_n.

1.1.2 Extension to discriminant analysis

In this section, we consider mixtures of Gaussians. Let $X^n = (X_1, \ldots, X_n)$ and $Y^n = (Y_1, \ldots, Y_n)$ be n i.i.d. samples where $X_i \sim \mathcal{N}(\mu_X^*, \Omega_X^{*-1})$ and $Y_i \sim \mathcal{N}(\mu_Y^*, \Omega_Y^{*-1})$. The authors generalize the previous theory to design a framework for BvM in discriminant analysis, which aims at predicting whether an independent new sample z is from the X-class or Y-class. For a given $(\mu_X, \mu_Y, \Omega_X, \Omega_Y)$, Fisher's QDA rule can be written as

$$\Delta(\mu_X, \mu_Y, \Omega_X, \Omega_Y) = -(z-\mu_X)^T\Omega_X(z-\mu_X) + (z-\mu_Y)^T\Omega_Y(z-\mu_Y) + \log(\frac{\det(X)}{\det(Y)}).$$

The aim of the authors is to find the asymptotic posterior distribution

$$\sqrt{n}V^{-1}(\Delta(\mu_X, \mu_Y, \Omega_X, \Omega_Y) - \hat{\Delta})|X^n, Y^n, z$$

with some appropriate variance V^2 and some prior distribution.

Linear discriminant analysis (LDA)

We assume $\Omega_X^* = \Omega_Y^*$, and therefore the QDA rule can be reduced to the LDA rule. For a given prior Π, the posterior distribution for LDA is given by

$$\Pi(B|X^n, Y^n) = \frac{\int_B \exp(l_n(\mu_X, \mu_Y, \Omega))d\Pi(\mu_X, \mu_Y, \Omega)}{\int \exp(l_n(\mu_X, \mu_Y, \Omega))d\Pi(\mu_X, \mu_Y, \Omega)},$$

where $l_n(\mu_X, \mu_Y, \Omega)$ is the log-likelihood function decomposed as $l_n(\mu_X, \mu_Y, \Omega) = l_X(\mu_X, \Omega) + l_Y(\mu_Y, \Omega)$ where $l_X(\mu_X, \Omega) = \frac{n}{2}\log\det(\Omega) - \frac{n}{2}\mathrm{tr}(\Omega\tilde{\Sigma}_X)$ with $\tilde{\Sigma}_X) = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_X)(X_i - \mu_X)^T$ and l_Y is defined similarly.

The LDA functional is here $\Delta(\mu_X, \mu_Y, \Omega_X, \Omega_Y) = -(z-\mu_X)^T\Omega_X(z-\mu_X) + (z-\mu_Y)^T\Omega_Y(z-\mu_Y)$. The authors define

$$\Phi = \frac{1}{2}\Omega^*\left((z-\mu_X^*)(z-\mu_X^*)^T - (z-\mu_Y^*)(z-\mu_Y^*)^T\right)\Omega^*,$$

$$\eta_X = 2(z-\mu_X^*), \quad \eta_Y = -2(z-\mu_Y^*),$$

$$\Omega_t = \Omega + \frac{2t}{\sqrt{n}}\Phi,$$

$$\mu_{X,t} = \mu_X + \frac{t}{\sqrt{n}}\eta_X, \quad \mu_{Y,t} = \mu_Y + \frac{t}{\sqrt{n}}\eta_Y,$$

$$\hat{\Sigma} = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})^T\right),$$

$$V^2 = 4||\Sigma^{*1/2}\Phi\Sigma^{*1/2}||_F^2 + \eta_X^T\Omega^*\eta_X + \eta_Y^T\Omega^*\eta_Y.$$

Assume that for some $\delta_n = o(1)$, A_n is a subset of

$$\left\{\sqrt{n}\left(||\mu_X - \mu_X^*||^2 + ||\mu_Y - \mu_Y^*||^2 + ||\Sigma - \Sigma^*||^2\right) \vee \sqrt{p}\left(||\mu_X - \mu_X^*|| + ||\mu_Y - \mu_Y^*|| + ||\Sigma - \Sigma^*||\right) \leq \delta_n\right\}.$$

The main result for LDA is the following:

Theorem 4.3 *Assume* $\max(||\Sigma^*||, ||\Omega^*||) = O(1)$, $p^2/n = o(1)$ and $V^{-1} = O(1)$. *If for a given prior* Π, *the following two conditions are satisfied:*

1. $\Pi(A_n|X^n, Y^n) = 1 - o_P(1)$,
2. *For any fixed* $t \in \mathbb{R}$, $\frac{\int_{A_n}\exp(l_n(\mu_{X,t},\mu_{Y,t},\Omega_t))d\Pi(\mu_X,\mu_Y,\Omega)}{\int_{A_n}\exp(l_n(\mu_X,\mu_Y,\Omega))d\Pi(\mu_X,\mu_Y,\Omega)} = 1 + o_P(1)$ *then*

$$\sup_{t\in\mathbb{R}}\left|\Pi\left(\sqrt{n}V^{-1}(\Delta(\mu_X,\mu_Y,\Omega) - \hat{\Delta}) \leq t|X^n\right) - P(Z \leq t)\right| = o_P(1)$$

,

where $Z \sim \mathcal{N}(0,1)$ and the centering is $\hat{\Delta} = \Delta(\bar{X}, \bar{Y}, \hat{\Sigma}^{-1})$.

One might wonder where the condition $V^{-1} = O(1)$ comes from. It is actually related to the separation of the two classes, because we can show that under the setting of the above theorem, if $||\mu_X^* - \mu_Y^*|| \geq c$ for some $c > 0$, then $V^{-1} = O(1)$.

Quadratic distriminant analysis (QDA)

In the general case where we do not have $\Omega_X^* = \Omega_Y^*$, the posterior distribution for QDA is defined as

$$\Pi(B|X^n, Y^n) = \frac{\int_B \exp(l_n(\mu_X,\mu_Y,\Omega_X,\Omega_Y))d\Pi(\mu_X,\mu_Y,\Omega_X,\Omega_Y)}{\int \exp(l_n(\mu_X,\mu_Y,\Omega_X,\Omega_Y))d\Pi(\mu_X,\mu_Y,\Omega_X,\Omega_Y)},$$

where $l_n(\mu_X,\mu_Y,\Omega_X,\Omega_Y) = l_X(\mu_X,\Omega_X) + l_Y(\mu_Y,\Omega_Y)$. The authors define

$$\Phi_X = -\Omega_X^*(\Sigma_X^* - (z-\mu_X^*)(z-\mu_X^*)^T)\Omega_X^*$$

$$\Phi_Y = -\Omega_Y^*(\Sigma_Y^* - (z-\mu_Y^*)(z-\mu_Y^*)^T)\Omega_Y^*$$

$$\eta_X = 2(z-\mu_X^*), \quad \eta_Y = -2(z-\mu_Y^*),$$

$$\Omega_{X,t} = \Omega_X + \frac{2t}{\sqrt{n}}\Phi_X, \quad \Omega_{Y,t} = \Omega_Y + \frac{2t}{\sqrt{n}}\Phi_Y$$

$$\mu_{X,t} = \mu_X + \frac{t}{\sqrt{n}}\eta_X, \quad \mu_{Y,t} = \mu_Y + \frac{t}{\sqrt{n}}\eta_Y,$$

$$\hat{\Sigma}_X = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T, \quad \hat{\Sigma}_Y = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})^T,$$

$$V^2 = 2||\Sigma_X^{*1/2}\Phi_X\Sigma_X^{*1/2}||_F^2 + 2||\Sigma_Y^{*1/2}\Phi_Y\Sigma_Y^{*1/2}||_F^2 + \eta_X^T\Omega^*\eta_X + \eta_Y^T\Omega^*\eta_Y.$$

Assume A_n is a subset of

$$\begin{aligned}\Big\{\sqrt{n}\left(||\mu_X-\mu_X^*||^2+||\mu_Y-\mu_Y^*||^2+||\Sigma-\Sigma^*||^2\right)\\ \vee\sqrt{p}\left(||\mu_X-\mu_X^*||+||\mu_Y-\mu_Y^*||+||\Sigma-\Sigma^*||\right)\\ \vee\sqrt{n/p}\left(||\Sigma_X-\Sigma_X^*||_F^2+||\Sigma_Y-\Sigma_Y^*||_F^2\right)\\ \vee\sqrt{p}\left(||\Sigma_X-\Sigma_X^*||_F+||\Sigma_Y-\Sigma_Y^*||_F\right)\leq\delta_n\end{aligned} \tag{1}$$

The main result for QDA is the following theorem.

Theorem 4 *Assume* $\max(||\Sigma^*||, ||\Omega^*||) = O(1)$, $p^3/n = o(1)$ and $V^{-1} = O(1)$. *If for a given prior* Π, *the following two conditions are satisfied:*

1. $\Pi(A_n|X^n, Y^n) = 1 - o_P(1)$,
2. *For any fixed* $t \in \mathbb{R}$, $\frac{\int_{A_n}\exp(l_n(\mu_{X,t},\mu_{Y,t},\Omega_{X,t},\Omega_{Y,t}))d\Pi(\mu_X,\mu_Y,\Omega_{X,}\Omega_{Y})}{\int_{A_n}\exp(l_n(\mu_X,\mu_Y,\Omega_X,\Omega_Y))d\Pi(\mu_X,\mu_Y,\Omega_X,\Omega_Y)} = 1 + o_P(1)$ *then*

$$\sup_{t\in\mathbb{R}}\left|\Pi\left(\sqrt{n}V^{-1}(\Delta(\mu_X,\mu_Y,\Omega_X,\Omega_Y)-\hat{\Delta})\leq t|X^n\right) - P(Z\leq t)\right| = o_P(1)$$

,

where $Z \sim \mathcal{N}(0,1)$ and the centering is $\hat{\Delta} = \Delta(\bar{X}, \bar{Y}, \hat{\Sigma}_X^{-1}, \hat{\Sigma}_Y^{-1})$.

1.2 Main steps of the proofs

In this section, we present the theoretical foundations of the results discussed above. A few preliminary results are discussed and the main ideas of the proofs are explained. We assume the conditions mentioned above to be true in this section.

For functionals of the covariance matrix

This result is obtained by showing that the moment-generating function of

$$\left(\frac{\sqrt{n}(\phi(\Sigma)-\phi(\widehat{\Sigma}))}{\sqrt{2}||\Sigma^{*1/2}\Phi\Sigma^{*1/2}||_F}\Big|X^n\right)$$

under the distribution $\Pi^{A_n}(.|X^n)$, given by: $\forall B, \Pi^{A_n}(B|X^n) = \frac{\Pi(A_n\cap B|X^n)}{\Pi(A_n|X^n)}$, converges in P-probability to the moment-generating function of Z. Lemma 1 allows such strategy.

Lemma 1 Let P_n be a probability kernel, and P be a probability measure. Assuming $\forall t \in \mathbb{R}, t \mapsto \int e^{tx}\mathrm{d}P(x) \in \mathbb{R}$ and $\int e^{tx}\mathrm{d}P_n(x) \xrightarrow{d} \int e^{tx}\mathrm{d}P(x)$, then the following equality holds:

$$\sup_{t\in\mathbb{R}}\left|P_n(]-\infty,t]) - P(]-\infty,t])\right| = o_P(1)$$

This convergence is proven by considering an expansion of the log-likelihood along Φ, which is the result of the following lemma:

Lemma 2 Assume $||\Sigma^*|| \vee ||\Omega^*|| = O(1)$. For any symmetric matrix Φ and the perturbed precision matrix defined earlier, for a set A_n satisfying our condition, we have $\forall\Omega \in A_n$:

$$l_n(\Omega_t) - l_n(\Omega) = \frac{t\sqrt{n}\mathrm{Tr}((\Sigma-\hat{\Sigma})\Phi)}{\sqrt{2}||\Sigma^{*1/2}\Phi\Sigma^{*1/2}||_F} - \frac{t^2||\Sigma^{1/2}\Phi\Sigma^{1/2}||_F}{2||\Sigma^{*1/2}\Phi\Sigma^{*1/2}||_F} - \frac{n}{2}\sum_{j=1}^{p}\int_0^{h_j}\frac{(hj-s)^2}{(1-s)^3}ds$$

Short version of the proof on functionals of the covariance matrix The moment-generating function f of

$$\left(\frac{\sqrt{n}(\phi(\Sigma)-\phi(\widehat{\Sigma}))}{\sqrt{2}||\Sigma^{*1/2}\Phi\Sigma^{*1/2}||_F}\Big|X^n\right)$$

under the distribution $\Pi^{A_n}(.|X^n)$ is computed as follows:

$$f(t)=\frac{\int_{A_n}\exp\Big(\frac{t\sqrt{n}(\phi(\Sigma)-\phi(\widehat{\Sigma}))}{\sqrt{2}||\Sigma^{*1/2}\Phi\Sigma^{*1/2}||_F}+l_n(\Omega)\Big)\mathrm{d}\Pi(\Omega)}{\int_{A_n}\exp(l_n(\Omega))\mathrm{d}\Pi(\Omega)}$$

Once this equality is derived, the objective is to approximate the following quantity:

$$\frac{t\sqrt{n}(\phi(\Sigma)-\phi(\widehat{\Sigma}))}{\sqrt{2}||\Sigma^{*1/2}\Phi\Sigma^{*1/2}||_F}+l_n(\Omega)$$

By $l_n(\Omega_t)$, which is done by :

- Applying the approximate-linearity assumption on $\mathrm{Tr}((\Sigma-\hat{\Sigma})\Phi)$
- Showing that $\frac{||\Sigma^{1/2}\Phi\Sigma^{1/2}||_F^2}{||\Sigma^{*1/2}\Phi\Sigma^{*1/2}||_F^2}\sim 1$ And $|\frac{n}{2}\sum_{1\leq j\leq p}\int_0^{h_j}\frac{(h_j-s)^2}{(1-s)^2}\mathrm{d}s|=o(1)$

Thus f becomes:

$$f=(1+o_P(1))\exp(\frac{t^2}{2})\underbrace{\frac{\int_{A_n}\exp(l_n(\Omega_t))\mathrm{d}\Pi(\Omega)}{\int_{A_n}\exp(l_n(\Omega))\mathrm{d}\Pi(\Omega)}}_{=1+o_P(1)\text{ by condition 2.}}=(1+o_P(1))\exp(\frac{t^2}{2})$$

$t\mapsto\exp(\frac{t^2}{2})$ being the moment generating function of $Z\sim\mathcal{N}(0,1)$, the result is proven.

The main idea of this proof is that the log of the likelihood ratio can be expanded on the sets A_n, this expansion is allowed by the hypothesis on the approximate-linearity of the functional. This expansion is then used to approximate the moment generating function of Π^{A_n} as "close to" the $\mathcal{N}(O,1)$ moment generating function.

The same approach can be applied to prove the theorem 2.2, with some small steps to link the two frameworks:

BvM for functional of the precision matrix Let $\psi: x\in\mathbb{R}^{p^2}\mapsto\psi(x)\in\mathbb{R}$ be some functional over the p^2-dimensional $\mathbb{R}$-space.
It is assumed that ϕ is approximately linear when approaching the truth parameter Ω^*, the approximate linearity assumption resumes to assume the existence of some symmetric matrix Ψ such as:

$$\sup_{A_n}\left[\left|\psi(\Omega)-\phi(\widehat{\Sigma}^{-1})-\mathrm{tr}((\Omega-\widehat{\Sigma}^{-1})\Psi)\right|\right]=o_P(\sqrt{n}||\Omega^{*1/2}\Psi\Omega^{*1/2}||_F)$$

On some set $A_n \subset \{\sqrt{rp}||\Sigma - \Sigma^*|| \leq \delta_n\}$ for some $\delta_n = o(1)$ and $r \geq \text{rank}(\Psi)$
With the same other assumptions than in the theorem above, and for the perturbed precision matrix:

$$\Omega_t = \Omega - \frac{\sqrt{2}t}{\sqrt{n}||\Omega^{*1/2}\Psi\Omega^{*1/2}||_F}\Omega^*\Psi\Omega^*$$

The following result holds:

$$\sup_{t\in\mathbb{R}}\left|\Pi\left(\frac{\sqrt{n}(\psi(\Omega) - \psi(\widehat{\Sigma}^{-1}))}{\sqrt{2}||\Omega^{*1/2}\Psi\Omega^{*1/2}||_F} \leq t\Big|X^n\right) - \mathbb{P}(Z \leq t)\right| \xrightarrow{P} 0$$

Where $Z \sim \mathcal{N}(0,1)$

To prove this result the authors only have to show that $|\psi(\Omega) - \psi(\widehat{\Omega}) - \text{tr}((\Sigma - \widehat{\Sigma})\Phi| = o_P(\sqrt{2/n}||\Omega^{*1/2}\Psi\Omega^{*1/2}||_F)$ to be able to use the expansion of Lemma 2 and follow the same steps as in the previous theorem.

Short version of the proof for functionals of the precision matrix

Let $\Phi = -\Omega^*\Psi\Omega^*$, which gives us:

$$||\Omega^{*1/2}\Psi\Omega^{*1/2}||_F = ||\Sigma^{*1/2}\Phi\Sigma^{*1/2}||_F$$

The expansion of $l_n(\Omega_t)$ on A_n becomes:

$$l_n(\Omega_t) - l_n(\Omega) = \frac{t\sqrt{n}\text{tr}((\Sigma - \widehat{\Sigma})\Phi)}{\sqrt{2}||\Omega^{*1/2}\Psi\Omega^{*1/2}||_F} - t^2/2 + o(1)$$

Then the idea is to approximate $\sqrt{n}\text{tr}((\Sigma - \widehat{\Sigma})\Phi)$ by $\sqrt{n}(\psi(\Omega) - \psi(\widehat{\Sigma}^{-1}))$.

The authors then derive the associated approximation error on A_n, we have for $V = 2||\Omega^{*1/2}\Psi\Omega^{*1/2}||_F^2$:

$$\begin{aligned}&\sqrt{n}V^{-1/2}|\psi(\Omega) - \psi(\widehat{\Omega}) - \text{tr}((\Sigma - \widehat{\Sigma})\Phi| \\ \leq &\sqrt{n}V^{-1/2}\Big(\underbrace{|\text{tr}((\Sigma - \widehat{\Sigma}\Omega^*\Psi(\Omega^* - \widehat{\Omega}))|}_{\mathcal{E}_1} + \underbrace{|\text{tr}((\Sigma - \widehat{\Sigma})(\Omega^* - \Omega)\Psi\widehat{\Omega})|}_{\mathcal{E}_2}\Big)\end{aligned}$$

Using the SVD of Ψ, we have for $d_1, \ldots, d_r$ the singular values of Ψ:

$$\mathcal{E}_1 = O_P(||\widehat{\Sigma} - \Sigma||.||\widehat{\Sigma} - \Sigma^*|| \sum_{1\leq l\leq r}|d_l|)$$

and:

$$\mathcal{E}_2 = O_P(||\widehat{\Sigma} - \Sigma||.||\widehat{\Sigma} - \Sigma^*|| \sum_{1\leq l\leq r}|d_l|)$$

Finally, we have: $\exists C > 0, V^{-1/2} \leq C||\Psi||_F^{-1}$ and:

$$\sqrt{n}V^{-1/2}|\psi(\Omega) - \psi(\widehat{\Omega}) - \text{tr}((\Sigma - \widehat{\Sigma})\Phi| = o_P(1)$$

This result gives us the expansion of l_n that is given before on A_n and the end of the proof follows the same steps as for the theorem on functionals of the covariance matrix.

1.3 Main points of the paper and links with classes

We saw in class that the classic approach was based on $P(\text{data}|\theta)$ and that the bayesian approach was based on $P(\theta|\text{data})$. The Bernstein-von Mises theorem establishes an important link between frequentist and bayesian approaches. In parametric models, the a posteriori distribution asymptotically concentrates around the parameter we wish to estimate regardless of the a priori distribution provided that we have enough observations. The theorem states that the centered a posteriori distribution is "asymptotically close" in probability to a normal law with variance the inverse of Fisher's information. This means that frequentist and bayesian approaches lead to similar results within parametric models, which supports the idea seen in class according to which those two approaches should be considered as complementary rather than opposed. This has interesting practical implications, such as the fact that a Bayesian 95%-confidence set must have frequentist coverage of about 95%, and conversely.

Furthermore, the maximum likelihood estimator is known to converge to a normal law with variance the inverse of Fisher's information, which means that the asymptotic law of the centered a posteriori distribution is the same as the asymptotic law of the MLE. This result is valuable since we can get a similar asymptotic distribution without the inconvenient of having to use the MLE. Indeed, we saw in class that the maximisation of the likelihood could be highly complex in some cases, and in particular in multidimensional and constraint configurations. We also saw that maximum likelihood estimators could be numerically unstable, which means that they can change a lot for little variations in the observations, especially for small sample sizes. Finally, this approach does not rely on probabilistic justifications and therefore its estimators do not provide sufficient material for decisional analysis - e.g. for tests.

Finally, we also noted in class how important was the covariance, which does not only control how close the realizations from the processes are to the mean function, but also controls other important properties such as smoothness. The extension provided by this article for the covariance functionals is therefore of great use, making it possible - among other things - to construct confidence intervals for the true values of the functional.

2 Application of the theorems to specific matrix functionals

2.1 Synthesis of the results obtained in the paper

The theorems above coupled with the lemmas allow to obtain **sufficient conditions** on the regime of (p, n) to obtain an asymptotic normality for a given functional. We synthesise this conditions in the table below[1]:

[1] We assume constant eigengap for LDA and QDA

Functional	$\phi(\hat{\Sigma})$ or $\psi(\hat{\Sigma}^{-1})$	Wishart prior (conjugate prior)	Gaussian prior (non-conj. prior)
σ_{ij}	No constraint	$p = o(n)$	$p^2 = o(n)$
ωij	$p^2 = o(n)$	$p^2 = o(n)$	$p^3 = o(n)$
$v^T \Sigma v$	No constraint	$p = o(n)$	$p^2 = o(n)$
$v^T \Omega v$	$p^2 = o(n)$	$p^2 = o(n)$	$p^3 = o(n)$
$\log\det(\Sigma)$	$p^3 = o(n)$	$p^3 = o(n)$	$p^3 = o(n)$
$\lambda_m(\Sigma)$	$p^2 = o(n)$	$p^2 = o(n)$	$p^4 = o(n)$
$\lambda_m(\Omega)$	$p^2 = o(n)$	$p^2 = o(n)$	$p^4 = o(n)$
LDA	$p^2 = o(n)$	$p^2 = o(n)$	$p^4 = o(n)$
QDA	$p^3 = o(n)$	$p^3 = o(n)$	$p^4 = o(n)$

We are going to explain how we obtain the first line of the table. The same reasoning can be applied for the other lines of the table.

Detailed explanation of line 1 : $\phi(\Sigma) = \sigma_{ij} = \mathbf{Tr}(\Sigma(\frac{E_{ij}+E_{ji}}{2}))$

- column 1 : The proof of the MLE is not provided in the paper. We provide our own proof.
  Let us define f_{ij} such that $f_{ij}(XX^T) = e_i^T XX^T e_j$.
  f_{ij} being linear, by applying the central limit theorem on $(Y_k = f_{ij}(X_k X_k^T))_k$, we get that $\sqrt{n}(\frac{\sum_{k=1}^n f_{ij}(X_k X_k^T)}{n} - E(f_{ij}(XX^T))) = \sqrt{n}(f_{ij}(\hat{\Sigma}) - f_{ij}(\Sigma)) \longrightarrow \mathcal{N}(0, Var(f_{ij}(XX^T)))$.
  $(X_i X_i^T)_i$ being i.i.d following a Wishart distribution $\mathcal{W}(\Sigma, 1)$, $Var(e_i^T XX^T e_j) = Var((XX^T)_{ij}) = (\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj})_{1\leq i,j\leq p}$ (the proof of the calculation of the variance can be found for example in [6] page 90).
  Hence : $Var(f_{ij}(XX^T)) = \sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}$.
  Hence, $\sqrt{n}(f_{ij}(\hat{\Sigma}) - f_{ij}(\Sigma)) = \sqrt{n}(\phi(\hat{\Sigma}) - \phi(\Sigma)) \longrightarrow \mathcal{N}(0, \sigma_{ij}^2 + \sigma_{ii}\sigma_{jj})$
  Note during this proof, no assumption had to be made on the regime of (p,n). Hence the "no constraint" in the table.

- column 2 : Since $\phi(\Sigma) - \phi(\hat{\Sigma}) = tr((\Sigma - \hat{\Sigma})(\frac{E_{ij}+E_{ji}}{2}))$, under the constraint that p = o(n), lemma 2.1 proves that the conditions of theorem 2.1 are satisfied.

- column 3 : Likewise since $\phi(\Sigma) - \phi(\hat{\Sigma}) = tr((\Sigma - \hat{\Sigma})(\frac{E_{ij}+E_{ji}}{2}))$, under the constraint that $p^2 = o(n/\log(n))$, lemma 2.2 proves that the conditions of theorem 2.1 are satisfied. By not considering the term log(n) ($\frac{n}{log(n)} \leq n$ for $n \geq 3$), we obtain that $p^2 = o(n)$

We have detailed an easy example here since the functional is linear in two senses :

- $\phi(\Sigma) - \phi(\hat{\Sigma}) = tr((\Sigma - \hat{\Sigma})(\frac{E_{ij}+E_{ji}}{2}))$
- $\phi(\Sigma + \lambda\tilde{\Sigma}) = \sigma_{ij} + \lambda\tilde{\sigma_{ij}}$

Indeed, by linearity (in the sense that $\phi(\Sigma) - \phi(\hat{\Sigma}) = tr((\Sigma - \hat{\Sigma})(\frac{E_{ij}+E_{ji}}{2}))$), we obviously do not need to find A_n satisfying $\sup_{A_n} \sqrt{n}||\Sigma^{*1/2}(\frac{E_{ij}+E_{ji}}{2})\Sigma^{*1/2}||_F^{-1}|\phi(\Sigma) - \phi(\hat{\Sigma}) - \mathrm{tr}((\Sigma - \hat{\Sigma})(\frac{E_{ij}+E_{ji}}{2}))| = o_P(1)$.
For the MLE, the fact that the functional is linear (in the sense that $\phi(\Sigma + \lambda\tilde{\Sigma}) = \sigma_{ij} + \lambda\tilde{\sigma_{ij}}$) makes the proof of $\phi(\hat{\Sigma})$ being the MLE easy and do not impose any constraint on (n,p). In a more general case, the delta method is needed.
For non linear functionals (in the sense that there exists Φ such that $\phi(\Sigma) - \phi(\hat{\Sigma}) = tr((\Sigma - \hat{\Sigma})\Phi)$) more work has to be done. We will not detail the calculation used to obtain the result of

the other lines of the tables as they are detailed in the paper. What is important to understand is the framework, detailed in this example, to obtain them.

2.2 Theoretical extension: another functional

In this section we are going to consider a new functional **that is not mentioned in the paper**:

$$\phi(\Sigma) = \mathrm{Tr}(\Sigma)$$

We claim the following results :

Proposition 1 :

1. $\mathrm{Tr}(\hat{\Sigma})$ is the MLE. No constraint is imposed on the regime of (p, n). And $\sqrt{n}(\phi(\hat{\Sigma}) - \phi(\Sigma)) \longrightarrow \mathcal{N}(0, 2||\Sigma||_F^2)$.

2. If $p = o(n)$ and $\max(||\Sigma^*||, ||\Omega^*||) = O(1)$, by considering the Wishart prior $\mathcal{W}(I, p+b-1)$ where b=O(1), then we have

$$P_{\Sigma^*}^n \left(\sup_{t\in\mathbb{R}} \left| \Pi \left(\frac{\sqrt{n}(\mathrm{Tr}(\Sigma) - \mathrm{Tr}(\hat{\Sigma})))}{\sqrt{2}||\Sigma^*||_F} \leq t | X^n \right) - P(Z \leq t) \right| \right) \to 0$$

3. If $p^2 = o(n)$ and $\max(||\Sigma^*||, ||\Omega^*||) = O(1)$, by considering the Gaussian prior, then we have

$$P_{\Sigma^*}^n \left(\sup_{t\in\mathbb{R}} \left| \Pi \left(\frac{\sqrt{n}(\mathrm{Tr}(\Sigma) - \mathrm{Tr}(\hat{\Sigma})))}{\sqrt{2}||\Sigma^*||_F} \leq t | X^n \right) - P(Z \leq t) \right| \right) \to 0$$

Proof :

1. Let $f(XX^T) = \mathrm{Tr}(XX^T) = \sum_{i=1}^p e_i^T XX^T e_i$.
   By applying the central limit theorem on $(Y_i = f(X_i X_i^T))_i$, we get by linearity of f that :

$$\sqrt{n}(f(\hat{\Sigma}) - f(\Sigma)) \longrightarrow \mathcal{N}(0, Var(f(XX^T)))$$

   $Var(f(XX^T)) = Var(\sum_{i=1}^p e_i^T XX^T e_i) = \sum_{i=1}^p Var(e_i^T XX^T e_i) + 2\sum_{i<j} cov(e_i^T XX^T e_i, e_j^T XX^T e_j)$.
   $(X_i X_i^T)_i$ being i.i.d following a Wishart distribution, in [6], page 90, it is proven that $cov(e_i^T XX^T e_i, e_j^T XX^T e_j) = 2\sigma_{ij}^2$ **(in the general case :** $cov(e_i^T XX^T e_j, e_k^T XX^T e_l) = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$**)**.

   Hence, $\sum_{i=1}^p Var(e_i^T XX^T e_i) + 2\sum_{i<j} cov(e_i^T XX^T e_i, e_j^T XX^T e_j) = 2\sum_{i=1}^p \sigma_{ii}^2 + 2\sum_{i<j} \sigma_{ij}^2 = 2||\Sigma||_F^2$. Hence the result.

2. $\phi(\sigma)$ is linear in the sense that $\phi(\Sigma) - \phi(\hat{\Sigma}) = tr((\Sigma - \hat{\Sigma})\Phi)$ with $\Phi = I_p$ the identity matrix. Therefore if $p = o(n)$ and $\max(||\Sigma^*||, ||\Omega^*||) = O(1)$, by lemma 2.1 and theorem 2.1, we obtain the result.

3. $\phi(\sigma)$ is linear in the sense that $\phi(\Sigma) - \phi(\hat{\Sigma}) = tr((\Sigma - \hat{\Sigma})\Phi)$ with $\Phi = I_p$ the identity matrix. Therefore if $p^2 = o(n)$ and $\max(||\Sigma^*||, ||\Omega^*||) = O(1)$, by lemma 2.2 and theorem 2.1, we obtain the result.

We obtain similar results for the precision matrix :

Proposition 2 :

1. If $p^2 = o(n)$, $\text{Tr}(\hat{\Omega})$ is the MLE. And $\sqrt{n}(\phi(\hat{\Omega}) - \phi(\Omega)) \longrightarrow \mathcal{N}(0, 2||\Omega||_F^2)$.

2. If $p^2 = o(n)$ and $\max(||\Sigma^*||, ||\Omega^*||) = O(1)$, by considering the Wishart prior $\mathcal{W}(I, p+b-1)$ where b=O(1), then we have

$$P_{\Omega^*}^n \left(\sup_{t \in \mathbb{R}} \left| \Pi \left(\frac{\sqrt{n}(\text{Tr}(\Omega) - \text{Tr}(\hat{\Sigma}^{-1}))}{\sqrt{2}||\Omega^*||_F} \leq t | X^n \right) - P(Z \leq t) \right| \right) \to 0$$

3. If $p^3 = o(n)$ and $\max(||\Omega^*||, ||\Omega^*||) = O(1)$, by considering the Gaussian prior, then we have

$$P_{\Omega^*}^n \left(\sup_{t \in \mathbb{R}} \left| \Pi \left(\frac{\sqrt{n}(\text{Tr}(\Omega) - \text{Tr}(\hat{\Sigma}^{-1}))}{\sqrt{2}||\Omega^*||_F} \leq t | X^n \right) - P(Z \leq t) \right| \right) \to 0$$

Proof : We will not give thorough details of the proofs as they are almost the same as in proposition 1. However we will highlight the keys differences.

1. The proof is the same except that we need to consider the inverse wishart distribution instead of the wishart distribution. The additional condition $p^2 = o(n)$ comes from the formula of the covariance related to the inverse wishart distribution where we can see that it is needed to obtain the convergence of the estimator.

2. Replace the use of theorem 2.1 by theorem 2.2. The stronger condition $p^2 = o(n)$ comes from the fact that $r = \text{rank}(\phi) = p$ in theorem 2.2.

3. Replace the use of theorem 2.1 by theorem 2.2. The stronger condition $p^3 = o(n)$ comes from the fact that $r = \text{rank}(\phi) = p$ in theorem 2.2.

3 Experimental protocol and results

In this section, we present some results of simulations on convergence rates of posterior distributions of functionals of covariance/precision matrices. We first present the setup we are using to extract those simulations and then the results.

3.1 Setup

At first, we wish to illustrate the results on the entry-wise functional on the precision matrix. We use a $\mathcal{W}_p(I, p+b-1)$ prior on Ω and we simulate the data $(X_1, \ldots, X_n) \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma^*)$ where $\Sigma^* \in \mathbb{R}^{p \times p}$ is some symmetric positive definite matrix.

Let f be a functional on $\mathbb{R}^{p \times p}$, we want to draw from the distribution $\Pi(\frac{\sqrt{n}(f(\Omega) - f(\hat{\Sigma}^{-1}))}{\sqrt{2}||\Omega^{*1/2}\Psi\Omega^{*1/2}||_F} | X^n)$ to compare it to $\mathcal{N}(0,1)$. For that the following procedure is retained:

1. Draw $A = [a_{i,j}]_{i,j}$ where $\forall i, j, a_{i,j} \sim \mathcal{U}(-1, 1)$, let $\Omega^* = A^\top A$ and $\Sigma^* = \Omega^{*^{-1}}$
2. Draw n i.i.d samples $(X_1, \ldots, X_n) \sim \mathcal{N}(0, \Sigma^*)$, let $\hat{\Sigma} = \frac{1}{n} \sum_{1 \leq i \leq n} X_i^\top X_i$ and $\hat{\Omega} = \hat{\Sigma}^{-1}$
3. Draw $(Y_1, \ldots, Y_N)$ i.i.d samples from $\Omega | X^n \sim \mathcal{W}_p((n\hat{\Sigma} + I)^{-1}, n + p + b - 1)$
4. Compute $\left\{ f_i = \frac{\sqrt{n}(f(\Omega_i) - f(\widehat{\Sigma}^{-1}))}{\sqrt{2}||\Omega^{*1/2} \Psi \Omega^{*1/2}||_F} \right\}_{i=1}^N$
5. Compute $\hat{F}(t) = \frac{1}{N} \sum_{1 \leq i \leq N} \mathbb{1}(f_i \leq t), \forall t \in \mathbb{R}$ where $F(t) = \Pi(\frac{\sqrt{n}(f(\Omega) - f(\hat{\Sigma}^{-1}))}{\sqrt{2}||\Omega^{*1/2} \Psi \Omega^{*1/2}||_F} \leq t | X^n)$
6. Estimate $\sup_{t \in \mathbb{R}} |F(t) - \mathcal{Z}(t)|$ by $\sup_{t \in \mathbb{R}} |\hat{F}(t) - \mathcal{Z}(t)|$ where $\mathcal{Z}$ is the cumulative density function of $\mathcal{N}(0, 1)$

First we compare the approximated distribution to the distribution of $Z \sim \mathcal{N}(0, 1)$ with $p = 10$ and increasing n. Then we make p increase and make $n(p)$ grow accordingly to the theoretical results presented above. Ultimately we compare the distributions of the MLE and $f(\Sigma)|X^n$ for different values of n.
The results below are obtained for three linear functionals, the entry-wise and the trace, and logdet which is nonlinear. The entry-wise functional is a functional on Ω and the other two are functionals on Σ. $N = 1000$, and Ω a random matrix with singular values in a restrained interval.
The confidence interval of level .95 for Monte-Carlo approximation on $F(t)$ is $F(t) \in]\hat{F}(t) - 5.10^{-4}, \hat{F}(t) + 5.10^{-4}[$.

3.2 Fixed parameters

We fix $p = 10$ and make n vary in $[1, 10000]$ for the following functionals: entry-wise, trace and log-determinant. The distances are averaged over 20 runs each.

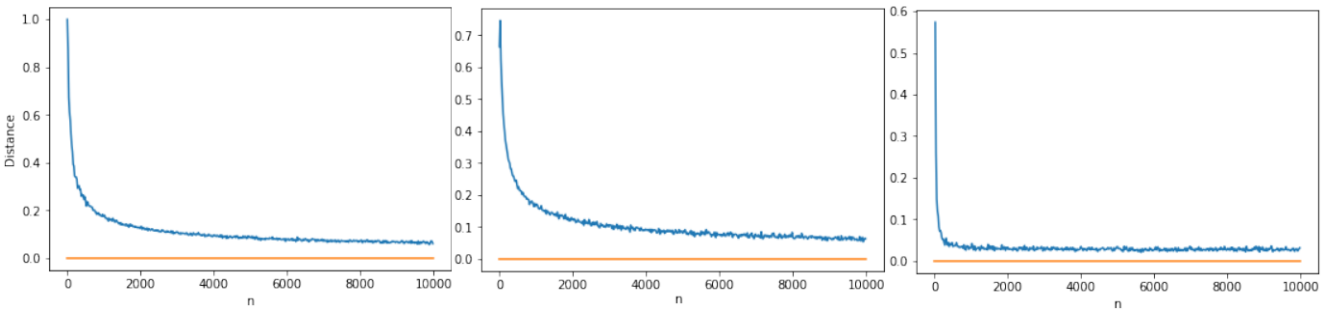

Figure 1: Evolutions for investigated functionals (entry-wise,trace,log-determinant) of $\sup_{t \in \mathbb{R}} |\widehat{F}(t) - \mathcal{Z}(t)|$

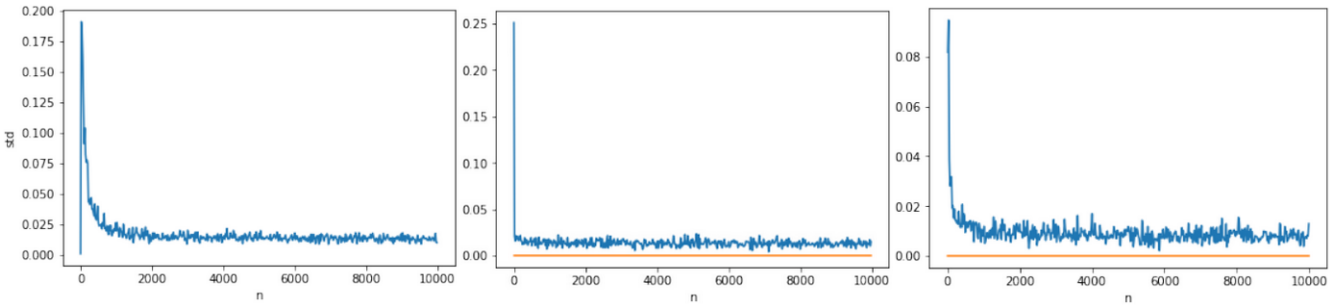

Figure 2: Evolutions for investigated functionals (entry-wise,trace,log-determinant) of $\widehat{\sigma}$

The distance between the distributions is indeed converging to 0. One point in analysing those results is that the standard deviation for 20 runs is very small, and even if the confidence intervals are drawn, they are not distinguishable from the mean here, and in the other results.

3.3 Increasing dimensionality

We now make p increase and compute n for multiple polynomial functions of p to compare the rate of convergence.

To achieve these results for Ω increasing in size one has to simulate random symmetric definite positive matrices of any size while controlling the scale of their eigenvalues. To do so, we draw the eigenvalues $\lambda_1, \ldots, \lambda_p$ from $\mathcal{U}(3,5)$, then draw a random unitary matrix P and compute $\Omega^* = P^\top \mathrm{Diag}(\lambda_1, \ldots, \lambda_p) P$. A new precision matrix is drawn this way for every (p, n) and every different run.

We run the simulation 20 times and average the errors, which gives us the graphs in Fig. 3.

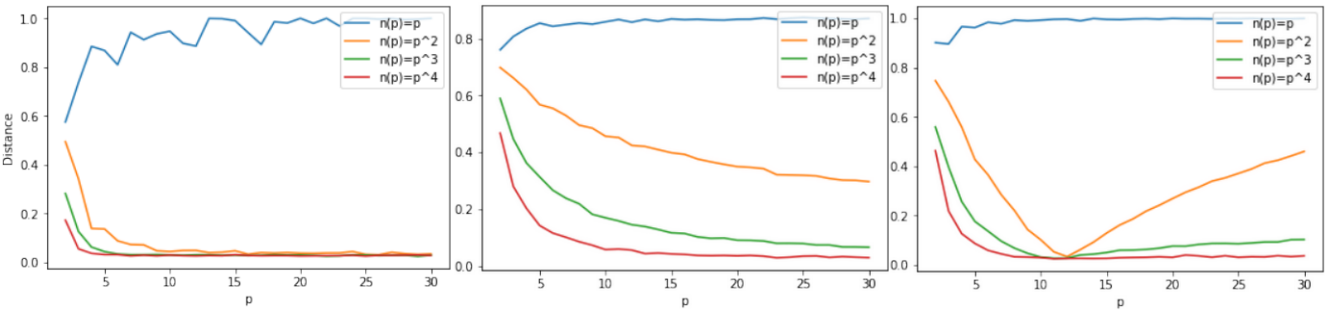

Figure 3: Distance to $\mathcal{N}(0,1)$, for investigated functionals(entry-wise,trace,log-determinant)

The hypotheses of convergence are $p^2/n = o(1)$ for the entry-wise functional, $p/n = o(1)$ for the trace and $p^3/n = o(1)$ for the log-determinant functional. Figure 3 shows that for $n(p)$ not satisfying those conditions, no sign of convergence is observed for trace and log-determinant, interestingly the distance for entry-wise functional of the precision matrix seems to converge even if $p^2/n = 1$. According to the authors the bounds for convergence are sharp but not necessarily optimal, the previous fact illustrates this statement.

3.4 Comparing frequentist and bayesian estimators

We fix p=10 then for each functionals, for a given true parameter we compare MLE and a posteriori distributions, with $n \in \{100, 10000\}$.

We fix Ω^* and draw N_s sets $(X_1^i, \dots, X_n^i)_{i=1}^{N_s} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \Omega^{*-1})$ which gives N_s MLE $(\hat{\Sigma}^i)_{i=1}^{N_s}$, then for each $i = 1, \dots, N_s$ we draw N_s samples from the posterior distribution of $\Omega|X^i$ and compute the distribution of $f(\Omega)|X^i$ the same way we did above. At the end we compare the overall distribution of $f(\Omega)|X$ and the distribution of $f(\hat{\Sigma})$.

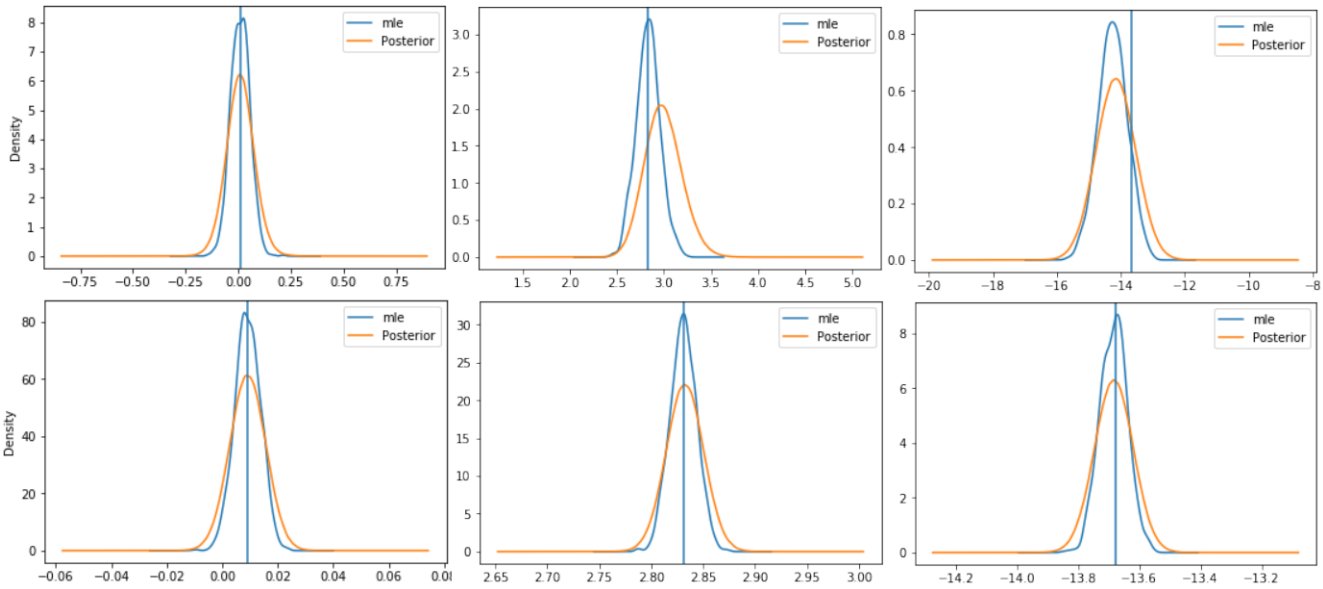

Figure 4: Density of MLE and a posteriori estimators for entry-wise, trace, log-determinant (from left to right), with $n = 100$ on the first line and $n = 10^4$ on the second one.

For the entry-wise functional, the mean of $f(\Omega)|X$ has converged to the mean of the MLE an $n = 100$ although it is not the case for the trace and log-determinant. For $n = 10^4$ all the convergence bounds are verified and all the distributions have the same mean. Only deviation of both reduces further. There is a gap between the variances of those distributions. This gap is explained by the variance of $f(\Omega) - f(\hat{\Omega})$ in the main theorems, which is different from the asymptotic variance of the MLE.

Concluding remarks

Summary

The BvM theorem establishes a valuable link between frequentist and bayesian approaches, and this article shows that it is possible to extend BvM results for matrix functionals when the dimension of the matrix p grows with the sample size n. The authors provide proofs for each of their theorems, showing that their theoretical approach is sound, and we presented the main points of these proofs. We proposed a theoretical extension of these results to the trace of the covariance matrix. We finally approximated the distribution of the posterior for three functionals and studied the differences with a standard normal law.

Limits of the results in infinite-dimensional models

As underlined above, one of the consequences of BvM theorem is that Bayesian confidence sets have good frequentist coverage properties and conversely. However, Freedman showed in 1965 [2] and Cox in 1993 [3] that this result does not hold if the random variable has an infinite countable probability space - even for the simplest infinite-dimensional models. This is due to the fact that according to Freedman [2], when sampling from a countably infinite population

with unknown distribution, for all but a set of priors of the first category, Bayes estimates are consistent only at a set of distributions of the first category. This means that for essentially all priors the Bayes estimates are consistent essentially nowhere.

Freedman provides examples to clarify this phenomenon in 1999 [4] that involve sequences of independent normal variables and rely on the following model: $Y_i = \beta_i + \epsilon_i$ for $i = 1, 2, \dots$, where the ϵ_i are iid normal random variables $\mathcal{N}(0, \sigma_n^2)$ with $\sigma_n \to 0$. He introduces n, that stands for the sample size, and says that for each n, the data consists of an infinite sequence $\{Y_{n,1}, Y_{n,2}, \dots\}$ with $Y_{n,i} = \beta_i + \epsilon_{n,i}$. The example focuses on one infinite-dimensional functional—the square of the l_2 norm: $T_n = \sum_{i=1}^{\infty}(\beta_i - \hat{\beta}_i)^2$. The Bayesian computes $\mathcal{L}(T_n|Y)$, while the frequentist computes $\mathcal{L}(T_n|\beta)$. It turns out that there is a radical difference between the asymptotic behavior of $\mathcal{L}(T_n|Y)$ and the asymptotic behavior of $\mathcal{L}(T_n|\beta)$ contrary to the finite-dimensional case, and that therefore the BvM theorem does not hold. This is due to two reasons:

• For the frequentist, the variance of $\hat{\beta}$ is driven by ϵ, and Freedman shows that this variance is smaller than the Bayes variance.

• The frequentist distribution of T_n is offset from the Bayesian distribution by arbitrarily large amounts, which is a consequence of "Bayes bias".

A consequence of that limit is that if frequentist coverage probabilities are wanted in an infinite-dimensional problem, then frequentist coverage probabilities must be computed. As for Bayesians, unless they are convinced of the fine details of their priors, they also need to proceed with caution in the infinite-dimensional case.

References

[1] Gao, C. and H. Zhou, (2016). Bernstein-von Mises theorems for functionals of the covariance matrix, Annals of Statistics, 10, 1751 – 1806.

[2] Freedman, David A. (1965). On the asymptotic behaviour of Bayes estimates in the discrete case II. The Annals of Mathematical Statistics, vol. 36, pp.454–456.

[3] Cox, D. (1993). An analysis of Bayesian inference for nonparametric regression. Annals of Statistics 21 903–923.

[4] Freedman, David A. (1999). On the Bernstein-von Mises Theorem with Infinite Dimensional Parameters.

[5] Cai, T. T., Liang, T. Zhou, H. H. (2013). Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional gaussian distributions. arXiv preprint arXiv:1309.0482 .

[6] Muirhead, R.J. (1982) Aspects of multivariate statistical theory, Wiley.