Bayesian Optimal Two-sample test for High-dimensional Gaussian Populations

Kyoungjae Lee, Kisung You, and Lizhen Lin (2023)

Presenter: Jae-Hoon Kim

May 8, 2024

Contents

1. Introduction

Data assumption

Consider two samples of observations from high-dimensional normal models

$$X_i \mid \mu_1, \Sigma_1 \overset{iid}{\sim} N_p(\mu_1, \Sigma_1), \quad i = 1, \ldots, n_1,$$
$$Y_i \mid \mu_2, \Sigma_2 \overset{iid}{\sim} N_p(\mu_2, \Sigma_2), \quad i = 1, \ldots, n_2,$$

where μ_1 & $\mu_2 \in \mathbb{R}^p$ are mean vectors and Σ_1 & $\Sigma_2 \in \mathbb{R}^{p \times p}$ are covariance matrices.

Two-sample tests

We are interested in testing the equality of mean vectors or covariance matrices.

Two-sample mean test

Assume $\Sigma_1 = \Sigma_2$ and test whether $\mu_1 = \mu_2$.

Two-sample covariance test

Given $\mu_1 = \mu_2 = 0$, test whether $\Sigma_1 = \Sigma_2$.

The types of two-sample tests

Existing tests can be divided into two types: l_2-type and max-type.

	l_2-type	max-type
Test statistics	Involves the l_2-type norm	Utilizes the maximum-type norm
Preferences	Many but small signals.	Few relatively large signals

In many applications, it is more natural to assume rare signals.

The examples of two-sample tests

Many Frequentist papers have proposed testing methods.

- Based on estimators of $\|A(\mu_1 - \mu_2)\|_2^2$ for some p.d matrix A. (Srivastava 2008)
- Based on estimators of $\|\Sigma_1 - \Sigma_2\|_F^2$. (Li and Chen 2012)
- Based on estimators of $tr(\Sigma_1^2)/\left\{tr(\Sigma_1)\right\}^2 - tr(\Sigma_2^2)/\left\{tr(\Sigma_2)\right\}^2$. (Srivastava 2010)
- Took the maximum of std. difference between sample covariances. (Cai 2013)

However, almost no theoretically supported Bayesian method has been proposed for high-dimensional two-sample tests.

2. Two-sample mean test

Data and Hypothesis

Suppose that we observe the data from two populations

$$X_i \mid \mu_1, \Sigma \overset{iid}{\sim} N_p(\mu_1, \Sigma), \quad i = 1, \dots, n_1,$$
$$Y_i \mid \mu_2, \Sigma \overset{iid}{\sim} N_p(\mu_2, \Sigma), \quad i = 1, \dots, n_2,$$

where μ_1 & $\mu_2 \in \mathbb{R}^p$ and Σ is a $p \times p$ covariance matrix.

Then, $X_{n_1} = (X_1, \dots, X_{n_1})^T \in \mathbb{R}^{n_1 \times p}$ and $Y_{n_2} = (Y_1, \dots, Y_{n_2})^T \in \mathbb{R}^{n_2 \times p}$ be the data matrices.

Data and Hypothesis

We are interested in the testing problem

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_1 : \mu_1 \neq \mu_2.$$

Bayesian hypothesis tests are typically based on Bayes factors.

$$BF = \frac{P[H_0|x] \times \pi_1}{P[H_1|x] \times \pi_0} = \frac{P[x|H_0]}{P[x|H_1]}.$$

Note that Bayes factor can be calculated in a closed form when $1 < p < n - 2$ in this case.

The maximum pairwise Bayes factor

To overcome this issue, we apply the maximum pairwise Bayes factor (mxPBF) approach.

Let $\tilde{X}_j = (X_{1j}, \ldots, X_{n_1 j})^T$ and $\tilde{Y}_j = (Y_{1j}, \ldots, Y_{n_1 j})^T$, for a given integer $1 \leq j \leq p$.

Then, we can marginalize the model as

$$\begin{aligned} \tilde{X}_j \mid \mu_{1j}, \sigma_{jj} &\overset{iid}{\sim} N_{n_1}(\mu_{1j} 1_{n_1}, \sigma_{jj} I_{n_1}), \\ \tilde{Y}_j \mid \mu_{2j}, \sigma_{jj} &\overset{iid}{\sim} N_{n_1}(\mu_{2j} 1_{n_2}, \sigma_{jj} I_{n_2}) \end{aligned}.$$

where $\mu_k = (\mu_{k1}, \ldots, \mu_{kp})^T$ for $k = 1, 2$, $\Sigma = (\sigma_{ij})$, and $1_q = (1, \ldots, 1)^T \in \mathbb{R}^q$.

The maximum pairwise Bayes factor

We can reformulate the testing problem

$$H_{0j} : \mu_{1j} = \mu_{2j} \quad vs \quad H_{1j} : \mu_{1j} \neq \mu_{2j},$$

in the sense that H_0 is true if and only if H_{0j} is true for all $j = 1, \ldots, p$.

By doing this, we can avoid inverse calculation and get Bayes factor even when $p \geq n - 2$ at the cost of ignoring the dependence structure of the data.

The maximum pairwise Bayes factor

mxPBF procedure

We first calculate pairwise Bayes factors(PBFs) based on $(\tilde{X}_j, \tilde{Y}_j)$ for $j = 1, \ldots, p$.

Priors under H_{0j}

$$\mu_j \mid \sigma_{jj} \sim N\left(\bar{Z}_j, \frac{\sigma_{jj}}{n\gamma}\right)$$

$$\pi\left(\sigma_{jj}\right) \propto \sigma_{jj}^{-1},$$

where $\mu_j = \mu_{1j} = \mu_{2j}$, $n = n_1 + n_2$, $\gamma = (n \vee p)^{-\alpha}$,
$\tilde{Z}_j = \left(\tilde{X}_j^T, \tilde{Y}_j^T\right)^T = (Z_{1j}, \ldots, Z_{nj})^T$, and $\bar{Z}_j = n^{-1} \sum_{i=1}^n Z_{ij}$.

The maximum pairwise Bayes factor

Priors under H_{1j}

$$\mu_{1j} \mid \sigma_{jj} \sim N\left(\bar{X}_j, \frac{\sigma_{jj}}{n_1 \gamma}\right),$$

$$\mu_{2j} \mid \sigma_{jj} \sim N\left(\bar{Y}_j, \frac{\sigma_{jj}}{n_2 \gamma}\right),$$

$$\pi\left(\sigma_{jj}\right) \propto \sigma_{jj}^{-1},$$

where $\bar{X}_j = {n_1}^{-1} \sum_{i=1}^{n_1} X_{ij},\ \bar{Y}_j = n_2^{-1} \sum_{i=1}^{n_2} Y_{ij}$

The maximum pairwise Bayes factor

Then, the resulting log PBF is

$$\log B_{10}\left(\tilde{X}_j, \tilde{Y}_j\right) := \log \frac{p\left(\tilde{X}_j, \tilde{Y}_j \mid H_{1j}\right)}{p\left(\tilde{X}_j, \tilde{Y}_j \mid H_{0j}\right)}$$

$$= \frac{1}{2} \log\left(\frac{\gamma}{1+\gamma}\right) + \frac{n}{2} \log\left(\frac{n\widehat{\sigma}^2_{Z_j}}{n_1\widehat{\sigma}^2_{X_j} + n_2\widehat{\sigma}^2_{Y_j}}\right)$$

where $H_v = v\left(v^T v\right)^{-1} v^T$ is the projection matrix, $\widehat{\sigma}^2_{Z_j} = n^{-1}\tilde{Z}_j^T\left(I_n - H_{1_n}\right)\tilde{Z}_j$, $\widehat{\sigma}^2_{X_j} = n_1^{-1}\tilde{X}_j^T\left(I_{n_1} - H_{1_{n_1}}\right)\tilde{X}_j$, and $\widehat{\sigma}^2_{Y_j} = n_2^{-1}\tilde{Y}_j^T\left(I_{n_2} - H_{1_{n_2}}\right)\tilde{Y}_j$.

Visualization of the approach

For a given integer $1 \leq j \leq p$,

$$X_{n_1} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_1,1} & \cdots & x_{n_1,j} & \cdots & x_{n_1,p} \end{bmatrix}, \quad Y_{n_2} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,j} & \cdots & y_{1,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{n_2,1} & \cdots & y_{n_2,j} & \cdots & y_{n_2,p} \end{bmatrix}$$

$$\overset{\textbf{Subtract}}{\xrightarrow{j^{th} column}} \quad \tilde{X}_j \mid \mu_{1j}, \sigma_{jj} = \begin{bmatrix} x_{1,j} \\ \vdots \\ x_{n_1,j} \end{bmatrix}, \quad \tilde{Y}_j \mid \mu_{2j}, \sigma_{jj} = \begin{bmatrix} y_{1,j} \\ \vdots \\ y_{n_2,j} \end{bmatrix} \quad \overset{\textbf{Calculate}}{\xrightarrow{PBF}} \quad B_{10}(\tilde{X}_j, \tilde{Y}_j).$$

The maximum pairwise Bayes factor

One can aggregate PBFs for all $j = 1, \ldots, p$ and define the mxPBF as

$$B^{\mu}_{\max,10}(X_{n_1}, Y_{n_2}) \quad := \quad \max_{1 \le j \le p} B_{10}(\tilde{X}_j, \tilde{Y}_j).$$

Thus, One can reject $H_0 : \mu_1 = \mu_2$ if the mxPBF is larger than a prespecified threshold.

Consistency of Bayes factor

Recall the testing problem.

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_1 : \mu_1 \neq \mu_2,$$

Theorem 2.1 says that the mxPBF is consistent under mild conditions.

Theorem 2.1

Assume $\log p \le n\epsilon_0$ and $\alpha > \frac{2(1+\epsilon_0)}{1-3\sqrt{C_1\epsilon_0}}$,

for some constant $0 < \epsilon_0 < 1$ and any constant $C_1 > 1$ arbitrarily close to 1.

Then, the mxPBF is consistent under H_0, i.e., $B^{\mu}_{\max,10}(X_{n_1}, Y_{n_2}) = O_p\left\{(n \vee p)^{-c}\right\}$,

for some constant $c > 0$.

Consistency of Bayes factor

Theorem 2.1 (Continued)

Assume there is at least one of indices $1 \le j \le p$ satisfying

$$\frac{n_1 n_2 \left(\mu_{01,j} - \mu_{02,j}\right)^2}{n^2 \sigma_{0,jj}} \ge \left[\sqrt{2C_1} + \sqrt{2C_1 + \alpha C_1 \left\{1 + (1 + 8C_1)\, \epsilon_0\right\}}\right]^2 \frac{\log(n \vee p)}{n}.$$

Then, the *mxPBF* is also consistent under H_1, *i.e.*,

$$\left\{B^{\mu}_{\max,10}\left(\mathbf{X}n_1, \mathbf{Y}n_2\right)\right\}^{-1} = O_p\left\{(n \vee p)^{-c'}\right\},$$ for some constant $c' > 0$.

Moreover, it is known that the condition under H_1 is minimax rate-optimal with respect to the maximum norm. (Cai et al. 2014)

3. Two-sample covariance test

Data and Hypothesis

Suppose that we observe the data from two populations

$$X_i \mid \Sigma_1 \overset{iid}{\sim} N_p(0, \Sigma_1), \quad i = 1, \ldots, n_1,$$
$$Y_i \mid \Sigma_2 \overset{iid}{\sim} N_p(0, \Sigma_2), \quad i = 1, \ldots, n_2,$$

where $\Sigma_1 = (\sigma_{1,ij}), \Sigma_2 = (\sigma_{2,ij})$ are $p \times p$ covariance matrices.

We encountered the testing problem

$$H_0 : \Sigma_1 = \Sigma_2 \quad vs \quad H_1 : \Sigma_1 \neq \Sigma_2.$$

The maximum pairwise Bayes factor

As in the mean test case, we need to divide the problem to apply

the maximum pairwise Bayes factor (mxPBF) approach.

Motivation for a division trick

Suppose the univariate random variables X and Y such that

$$X, Y \mid \mu, \Sigma \sim N_2(\mu, \Sigma), \ \textit{where } \mu = (\mu_x, \mu_y)^T \textit{ and } \Sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix}.$$

Then, we can see that $Y \mid X = x \sim N(\mu_y - \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \ \sigma_{yy}(1 - \rho^2))$,

where $\sqrt{\sigma_{yy}} = \sigma_y$, $\sqrt{\sigma_{xx}} = \sigma_x$, and $\rho = \sigma_{xy}/\sigma_x\sigma_y$.

The maximum pairwise Bayes factor

For a given pair (i, j) $1 \leq i \neq j \leq p$, we can induce the conditional distributions as

$$\tilde{X}_i \mid \tilde{X}_j, a_{1,ij}, \tau_{1,ij} \overset{iid}{\sim} N_{n_1}(a_{1,ij}\tilde{X}_j, \tau_{1,ij} I_{n_1}),$$
$$\tilde{Y}_i \mid \tilde{Y}_j, a_{2,ij}, \tau_{2,ij} \overset{iid}{\sim} N_{n_2}(a_{2,ij}\tilde{Y}_j, \tau_{2,ij} I_{n_2})$$

where $a_{k,ij} = \sigma_{k,ij}/\sigma_{k,jj}$, $\tau_{k,ij} = \sigma_{k,ii}(1 - \rho^2_{k.ij})$, and $\rho_{k,ij} = \sigma_{k,ij}/(\sigma_{k,ii}\sigma_{k,jj})^{1/2}$, $k = 1, 2$.

Then, we can reformulate the testing problem

$$H_{0,ij} : a_{1,ij} = a_{2,ij} \textit{ and } \tau_{1,ij} = \tau_{2,ij} \quad vs \quad H_{1,ij} : \textit{not } H_{0,ij},$$

in the sense that H_0 is true if and only if $H_{0,ij}$ is true for all pairs $1 \leq i \neq j \leq p$.

The maximum pairwise Bayes factor

Similarly, we first calculate pairwise Bayes factors(PBFs) based on $(\tilde{X}_i, \tilde{Y}_i, \tilde{X}_j, \tilde{Y}_j)$.

Priors under $H_{0,ij}$

$$a_{ij} \mid \tau_{ij} \sim N\left(\widehat{a}_{ij}, \frac{\tau_{ij}}{\gamma \|\tilde{Z}_j\|_2^2}\right),$$

$$\tau_{ij} \sim IG\left(a_0, b_{0,ij}\right),$$

where $a_0, b_{0,ij}$ are positive constants, $a_{ij} = a_{1,ij} = a_{2,ij}$, $\tau_{ij} = \tau_{1,ij} = \tau_{2,ij}.$, $\gamma = (n \vee p)^{-\alpha}$, and $\widehat{a}_{ij} = \tilde{Z}_i^T \tilde{Z}_j / \|\tilde{Z}j\|_2^2$.

The maximum pairwise Bayes factor

Priors under $H_{1,ij}$

$$a_{1,ij} \mid \tau_{1,ij} \sim N\left(\widehat{a}_{1,ij}, \frac{\tau_{1,ij}}{\gamma\|\tilde{X}_j\|_2^2}\right), \quad a_{2,ij} \,\Bigg|\, \tau_{2,ij} \sim N\left(\widehat{a}_{2,ij}, \frac{\tau_{2,ij}}{\gamma\|\tilde{Y}_j\|_2^2}\right),$$

$$\tau_{1,ij} \sim IG\left(a_0, b_{01,ij}\right), \quad \tau_{2,ij} \sim IG\left(a_0, b_{02,ij}\right),$$

where $b_{01,ij}$ and $b_{02,ij}$ are positive constants, $\widehat{a}_{1,ij} = \tilde{X}_i^T \tilde{X}_j / \|\tilde{X}_j\|_2^2$, and $\widehat{a}_{2,ij} = \tilde{Y}_i^T \tilde{Y}_j / \|\tilde{Y}_j\|_2^2$.

The maximum pairwise Bayes factor

Then, the resulting log PBF is

$$
\begin{aligned}
\log B_{10}\left(\tilde{X}_i, \tilde{Y}_i, \tilde{X}_j, \tilde{Y}_j\right) &:= \log \frac{p\left(\tilde{X}_i, \tilde{Y}_i \mid \tilde{X}_j, \tilde{Y}_j, H_{1,ij}\right)}{p\left(\tilde{X}_i, \tilde{Y}_i \mid \tilde{X}_j, \tilde{Y}_j, H_{0,ij}\right)} \\
&= \frac{1}{2}\log\left(\frac{\gamma}{1+\gamma}\right) + \log\Gamma\left(\frac{n_1}{2}+a_0\right) + \log\Gamma\left(\frac{n_2}{2}+a_0\right) \\
&- \log\Gamma\left(\frac{n}{2}+a_0\right) + \log\left(\frac{b_{01,ij}^{a_0} b_{02,ij}^{a_0}}{b_{0,ij}^{a_0}\Gamma\left(a_0\right)}\right) - \left(\frac{n_1}{2}+a_0\right)\log\left(b_{01,ij}+\frac{n_1}{2}\widehat{\tau}_{1,ij}\right) \\
&- \left(\frac{n_2}{2}+a_0\right)\log\left(b_{02,ij}+\frac{n_2}{2}\widehat{\tau}_{2,ij}\right) + \left(\frac{n}{2}+a_0\right)\log\left(b_{0,ij}+\frac{n}{2}\widehat{\tau}_{ij}\right),
\end{aligned}
$$

The maximum pairwise Bayes factor

where Γ is the Gamma function, $\widehat{\tau}_{ij} = n^{-1} \tilde{Z}_i^T \left(I_n - H_{\tilde{Z}_j} \right) \tilde{Z}_i$, $\widehat{\tau}_{1,ij} = n_1^{-1} \tilde{X}_i^T \left(I_{n_1} - H_{\tilde{X}_j} \right) \tilde{X}_i$, and $\widehat{\tau}_{2,ij} = n_2^{-1} \tilde{Y}_i^T \left(I_{n_2} - H_{\tilde{Y}_j} \right) \tilde{Y}_i$.

For a given threshold C_{cp}, one can reject $H_0 : \Sigma_1 = \Sigma_2$ if

$$B^{\Sigma}_{\max,10} \left(\mathbf{X}_{n_1}, \mathbf{Y}_{n_2} \right) := \max_{i \neq j} B_{10} \left(\tilde{X}_i, \tilde{Y}_i, \tilde{X}_j, \tilde{Y}_j \right) > C_{cp}.$$

Note that this approach treats each i^{th} and j^{th} variables as they are independent of the other p-2 variables.

Visualization of the approach

For a given pair (i, j) $1 \leq i \neq j \leq p$,

$$X_{n_1} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,i} & x_{1,j} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{n_1,1} & \cdots & x_{n_1,i} & x_{n_1,j} & \cdots & x_{n_1,p} \end{bmatrix}, \quad Y_{n_2} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,i} & y_{1,j} & \cdots & y_{1,p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y_{n_2,1} & \cdots & y_{n_2,i} & y_{n_2,j} & \cdots & y_{n_2,p} \end{bmatrix}$$

$$\xrightarrow[]{\substack{\textbf{Subtract}\\ i^{th} \& j^{th} column}} \tilde{X}_i \mid \tilde{X}_j, a_{1,ij}, \tau_{1,ij} = \begin{bmatrix} x_{1,i} \\ \vdots \\ x_{n_1,i} \end{bmatrix}, \tilde{Y}_i \mid \tilde{Y}_j, a_{2,ij}, \tau_{2,ij} = \begin{bmatrix} y_{1,i} \\ \vdots \\ y_{n_2,i} \end{bmatrix} \xrightarrow[]{\substack{\textbf{Calculate}\\ PBF}} B_{10}\left(\tilde{X}_i, \tilde{Y}_i, \tilde{X}_j, \tilde{Y}_j\right)$$

Consistency of Bayes factor

To show mxPBF is consistent, we will introduce some conditions.

Condition A1 (Assumption about high dimension)

$\epsilon_{0k} := \log(n \vee p)/n_k = o(1)$, for $k = 1, 2$,

which roughly means $p = O(\exp(n^c))$, for $0 < c < 1$.

Condition A2 (Assumption about correlation under H_0)

$\min_{i \neq j} \tau_{0,ij} \gg \{\log(n \vee p)\}^{-1}$, where $\tau_{0,ij} = \sigma_{0,ii}(1 - \rho_{0,ij}^2)$ is from true covariance Σ_0.

This is satisfied if $\min_{1 \leq i \leq p} \sigma_{0,ii} > \epsilon$ and $\max_{i \neq j} \rho_{0,ij}^2 < 1 - \epsilon$ for $\epsilon > 0$.

However, it allows $\sigma_{0,ii} \to 0$ and $\rho_{0,ij}^2 \to 1$ as $p \to \infty$ at certain rates.

Consistency of Bayes factor

Under H_0, the above condition $(A2)$ is a sufficient condition for consistency.

Under H_1, we assume that $(\Sigma_{01}, \Sigma_{02})$ satisfies the following condition $(A3)$ or $(A3^\star)$.

Condition A3 (Assumption about correlation under $\mathbf{H_1}$)

$\exists(i,j)$ with $i \neq j$ $s.t.$ $\{\log(n \vee p)\}^{-1} \ll \tau_{01,ij} \wedge \tau_{02,ij} \leq \tau_{01,ij} \vee \tau_{02,ij} \ll (n \vee p)$ satisfying either $\frac{\tau_{01,ij}}{\tau_{02,ij}} > \frac{1+C_{\mathrm{bm}}\sqrt{\epsilon_{01}}}{1-4\sqrt{C_1(\epsilon_{01}\vee\epsilon_{02})}}$ or $\frac{\tau_{02,ij}}{\tau_{01,ij}} > \frac{1+C_{\mathrm{bm}}\sqrt{\epsilon_{02}}}{1-4\sqrt{C_1(\epsilon_{01}\vee\epsilon_{02})}}$, for $C_{\mathrm{bm}}^2 > 8(\alpha+1)$ and $C_1 > 1$ arbitrarily close to 1 and $\tau_{0k,ij} = \sigma_{0k,ii}(1-\rho_{0k,ij}^2)$ is from true covariances $\Sigma_1 \neq \Sigma_2$.

Consistency of Bayes factor

Condition A3* (Assumption about varainces under H_1)

$\exists(i,j)$ with $i \neq j$ $s.t$ $\sigma_{01,ii} \vee \sigma_{02,ii} \ll (n \vee p)$ and

$$(a_{01,ij} - a_{02,ij})^2 \geq \frac{25}{2} C_1 \sum_{k=1}^{2} \left\{ \frac{\tau_{0k,ij}\epsilon_{0k}}{\sigma_{0k,jj}\left(1 - 2\sqrt{C_1 \epsilon_{0k}}\right)} \right\},$$

$$(a_{01,ij} - a_{02,ij})^2 \geq \frac{10n}{n + 2a_0} \sum_{k=1}^{2} \left\{ \frac{\epsilon_{0k}}{\sigma_{0k,jj}\left(1 - 2\sqrt{C_1 \epsilon_{0k}}\right)} \right\}$$

$$\times \left[\frac{b_{0,ij}}{\log(n \vee p)} + \left\{ \sum_{k=1}^{2} \sigma_{0k,ii}\left(1 + 4\sqrt{C_1 \epsilon_{0k}}\right) + \frac{2b_{0,ij}}{n} \right\} C_{\mathrm{bm},a} \right]$$

for $C_{\mathrm{bm},a} > \alpha + a_0 + 1$, $C_1 > 1$ arbitrarily close to 1, and $a_{k,ij} = \sigma_{k,ij}/\sigma_{k,jj}$.

Consistency of Bayes factor

Note that condition $(A3)$ or $(A3^\star)$ can be transformed into simpler conditions.

Define a class of two matrices

$H_1(C_{\mathrm{bm}}, C_{\mathrm{bm},a}) := \{(\Sigma_1, \Sigma_2) : (\Sigma_1, \Sigma_2)$ satisfies condition (A3) or $(A3^\star)\}$.

Condition (3.11)

Suppose that

$$\max_{1\le k\le 2} \max_{1\le i\ne j\le p} \rho_{0k,ij}^2 \le 1 - c_0,$$

$$\{\log(n \vee p)\}^{-1} \ll \min_{1\le k\le 2} \min_{1\le i\le p} \sigma_{0k,ii} \le \max_{1\le k\le 2} \max_{1\le i\le p} \sigma_{0k,ii} \ll (n \vee p),$$

for some small constant $c_0 > 0$.

Consistency of Bayes factor

Note that simplified condition characterizes the difference between Σ_{01} and Σ_{02} using the squared maximum standardized difference.

Simplified condition of (A3) or (A3*)

If $\alpha > 1$, $n_1 \asymp n_2$ and

$$\widetilde{H}1\left(C_\star, c_0\right) := \Big\{(\Sigma_1, \Sigma_2) : \max_{1\le i\le j\le p} \frac{(\sigma 1,ij - \sigma 2,ij)^2}{\sigma 1,ii\sigma 1,jj + \sigma_{2,ii}\sigma_{2,jj}} \ge C_\star \frac{\log(n \vee p)}{n},$$

$$\left(\Sigma_1, \Sigma_2\right) \text{ satisfies condition (3.11) with } c_0\Big\},$$

then $\widetilde{H}1\left(C_\star, c_0\right) \subset H_1\left(C_{\rm bm}, C_{{\rm bm},a}\right)$ for some large constant $C_\star > 0$.

Consistency of Bayes factor

Consider the two-sample covariance test $H_0 : \Sigma_1 = \Sigma_2$ vs $H_1 : \Sigma_1 \neq \Sigma_2$.

Note that the condition $\lim_{(n_1 \wedge n_2) \to \infty} n_1/n = 1/2$ can be relaxed to $n_1 \asymp n_2$.

Theorem 3.1

Assume $\lim_{(n_1 \wedge n_2) \to \infty} n_1/n = 1/2$ and condition (A1) holds.

Then, under H_0, if $\alpha > 12$ and condition (A2) holds, for some constant $c > 0$,

$B\max, 10^{\Sigma}\left(\mathbf{X}n_1, \mathbf{Y}_{n_2}\right) = O_p\left\{(n \vee p)^{-c}\right\}.$

Under H_1, if $(\Sigma_{01}, \Sigma_{02}) \in H_1\left(C_{\mathrm{bm}}, C_{\mathrm{bm},a}\right)$, for some constant $c' > 0$,

$\left\{B^{\Sigma}_{\max,10}\left(\mathbf{X}n_1, \mathbf{Y}n_2\right)\right\}^{-1} = O_p\left\{(n \vee p)^{-c}\right\}.$

Consistency of Bayes factor

Furthermore, condition $(A3)$ or $(A3^{\star})$ is rate optimal to guarantee consistency under H_0 as well as H_1.

Theorem 3.2

Let $\mathbb{E}_{\Sigma_{01},\Sigma_{02}}$ be the expectation with $(\Sigma_{01}, \Sigma_{02})$ and let $0 < \beta_0 < 1$.

Suppose $n_1 \asymp n_2$ and $p \geq n^c$ for some constant $c > 0$. Then, $\exists C_1, C_{bm}$ and $C_{bm,a} > 0$

$s.t.$ for large n, $\inf_{\phi \in \mathcal{T}} \sup_{(\Sigma_{01},\Sigma_{02}) \in H_1(C_{bm}, C_{bm,a})} \mathbb{E}_{\Sigma_{01},\Sigma_{02}}(1 - \phi) \geq \beta_0$,

where $\mathcal{T}$ is the set of tests over the multivariate nuomal $dist^n$ $s.t.$ $\mathbb{E}_0 \phi \to 0$ as $n \to \infty$ for any $\phi \in \mathcal{T}$ and $\mathbb{E}_0$ is the expectation under H_0.

4. Numerical results

Choice of hyperparameters

The proposed mxPBF approach requires hyperparameters

- α for mean vectors
- $a_0, b_{0,ij}, b_{01,ij}, b_{02,ij}$ and α for covariance matrices.

We suggest using $a_0 = b_{0,ij} = b_{01,ij} = b_{02,ij} = 0.01$ for all $1 \leq i \neq j \leq p$, since the above choice of hyperparameters does not affect the mxPBF too much.

Choice of α

- However, the choice of α in the mxPBFs is crucial to the performance.
- Choosing α according to theorems 2.1 and 3.1 might be overly conservative in practice.
- Therefore, we suggest to choose α that controls an empirical false positive rate (FPR) at a prespecified level.

Empirical FPR method

Suppose we have samples from two populations X_{n_1} and Y_{n_2}.

- Get the pooled sample mean vector and covariance matrix from samples
  - $\hat{\mu}_{\text{pool}} = (n_1\hat{\mu}_1 + n_2\hat{\mu}_2)/n$, where $\hat{\mu}_1 = \sum_{i=1}^{n_1} X_i/n_1$ and $\hat{\mu}_2 = \sum_{i=1}^{n_2} Y_i/n_2$.
  - $\hat{\Sigma}_{\text{pool}} = \left\{\sum_{i=1}^{n_1}(X_i - \hat{\mu}_1)(X_i - \hat{\mu}_1)^T + \sum_{i=1}^{n_2}(Y_i - \hat{\mu}_2)(Y_i - \hat{\mu}_2)^T\right\}/(n-2)$
- If $\hat{\Sigma}_{\text{pool}}$ is not positive definite, we add $\left\{-\lambda_{\min}\left(\hat{\Sigma}_{\text{pool}}\right) + 0.1^3\right\} I_p$ to $\hat{\Sigma}_{\text{pool}}$.
- Generate a simulated dataset $\mathbf{X}_{\text{sim}} = (X_{1,\text{ sim}}, \ldots, X_{n_1,\text{ sim}})^T \in \mathbb{R}^{n_1 \times p}$ and $\mathbf{Y}_{\text{sim}} = (Y_{1,\text{ sim}}, \ldots, Y_{n_2,\text{ sim}})^T \in \mathbb{R}^{n_2 \times p}$, where $X_{i,\text{ sim}}\mathrm{s}$ and $Y_{i,\text{ sim}}\mathrm{s}$ are random samples from $N_p\left(\hat{\mu}_{\text{pool}}, \hat{\Sigma}_{\text{pool}}\right)$.

Empirical FPR method

- By generating N simulated datasets $\left(\mathbf{X}_{\text{sim}}^{(s)}, \mathbf{Y}_{\text{sim}}^{(s)}\right)_{s=1}^{N}$, we can calculate the following empirical FPR for each α,

$$\widehat{\mathrm{FPR}}_{\alpha} = N^{-1} \sum_{s=1}^{N} I\left(B_{\max,10,\alpha}\left(\mathbf{X}_{\text{sim}}^{(s)}, \mathbf{Y}_{\text{sim}}^{(s)}\right) > 10\right).$$

- Among a grid of values, $\alpha \in \{0.01, 0.02, \ldots, 15\}$ for example, we can select the minimum value of α that achieves a prespecified FPR level.

Visualization of empirical FPR method

$$X_{n_1} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n_1,1} & \cdots & x_{n_1,p} \end{bmatrix}, \quad Y_{n_2} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,p} \\ \vdots & \ddots & \vdots \\ y_{n_2,1} & \cdots & y_{n_2,p} \end{bmatrix} \quad \xrightarrow{\textit{Pooling}} \quad (\hat{\mu}_{\text{pool}}, \hat{\Sigma}_{\text{pool}})$$

$$\xrightarrow{\textit{Generating}} \quad \mathbf{X}^{(s)}_{\text{sim}} = \begin{bmatrix} x^{(s)}_{1,1} & \cdots & x^{(s)}_{1,p} \\ \vdots & \ddots & \vdots \\ x^{(s)}_{n_1,1} & \cdots & x^{(s)}_{n_1,p} \end{bmatrix}, \quad \mathbf{Y}^{(s)}_{\text{sim}} = \begin{bmatrix} y^{(s)}_{1,1} & \cdots & y^{(s)}_{1,p} \\ \vdots & \ddots & \vdots \\ y^{(s)}_{n_2,1} & \cdots & y^{(s)}_{n_2,p} \end{bmatrix}$$

Visualization of empirical FPR method

For grid $\alpha = \{\alpha_1 = 0.01, \alpha_2 = 0.02, \ldots, \alpha_{n_\alpha} = 15\}$ and $1 \le s \le N$, calculate mxPBF

$$B_{\max,10,\alpha} = \begin{bmatrix} B_{\max,10,\alpha_1}\left(\mathbf{X}_{\text{sim}}^{(1)}, \mathbf{Y}_{\text{sim}}^{(1)}\right) & \cdots & B_{\max,10,\alpha_1}\left(\mathbf{X}_{\text{sim}}^{(N)}, \mathbf{Y}_{\text{sim}}^{(N)}\right) \\ \vdots & \ddots & \vdots \\ B_{\max,10,\alpha_{n_\alpha}}\left(\mathbf{X}_{\text{sim}}^{(1)}, \mathbf{Y}_{\text{sim}}^{(1)}\right) & \cdots & B_{\max,10,\alpha_{n_\alpha}}\left(\mathbf{X}_{\text{sim}}^{(N)}, \mathbf{Y}_{\text{sim}}^{(N)}\right) \end{bmatrix}$$

$$\xrightarrow[]{\substack{\textbf{Conditional} \\ \textit{rowSum}}} \quad \widehat{\mathrm{FPR}}_\alpha = \begin{bmatrix} N^{-1}\sum_{s=1}^{N} I\left(B_{\max,10,\alpha_1}\left(\mathbf{X}_{\text{sim}}^{(s)}, \mathbf{Y}_{\text{sim}}^{(s)}\right) > 10\right) \\ \vdots \\ N^{-1}\sum_{s=1}^{N} I\left(B_{\max,10,\alpha_{n_\alpha}}\left(\mathbf{X}_{\text{sim}}^{(s)}, \mathbf{Y}_{\text{sim}}^{(s)}\right) > 10\right) \end{bmatrix}$$

Visualization of empirical FPR method

When prespecified FPR is given as 0.05, we can choose $\hat{\alpha}$ that yields empirical rate closest to that value, i.e.,

$\hat{\alpha} = \alpha_k$ s.t. $\arg\min_{\alpha_k} abs \left\{ N^{-1} \sum_{s=1}^{N} I \left(B_{\max,10,\alpha_k} \left(\mathbf{X}_{\text{sim}}^{(s)}, \mathbf{Y}_{\text{sim}}^{(s)} \right) > 10 \right) - 0.05 \right\}$

Real data analysis

Datasets

- Small round blue cell tumors (SRBCT)
  - 11 cases of Burkitt lymphoma(BL) ($n_1 = 11$)
  - 18 cases of newroblastoma(NB) ($n_2 = 18$)
  - Data with 2308 genes ($p = 2308$)
- Prostate cancer
  - 52 patients with prostate tumors ($n_1 = 52$)
  - 50 patients with normal prostate ($n_2 = 50$)
  - Select 5000 genes with the largest absolute values of the t-statistics. ($p = 5000$)

Real data analysis

Contenders

- BS: l_2-type test proposed by Bai and Saranadasa (1996).
- SD: l_2-type test proposed by Srivastava and Du (2008).
- Sch: l_2-type test proposed by Schott (2007).
- LC: l_2-type test proposed by Li and Chen (2012).
- CLX: max-type test proposed by Cai et al. (2014)

  based on the constrained l_1-minimization for inverse matrix estimation (CLIME).
- CLX.AT: max-type test proposed by Cai et al. (2014)

  based on the inverse of the adaptive thresholding estimator.

Summary

- We introduce a Bayesian two-sample mean and covariance test in high-dimensional settings based on the idea of maximum pairwise Bayes factor (Lee et al. 2021).
- Tests are computationally scalable and consistent under relatively weak conditions compared to existing tests.
- We can choose α utilizing empirical FPR method.
- Proposed test shows clear advantages over other state-of-the-art methods in various scenarios.

Q&A