Biometrika (2024), **104**, 1, *p.* 1
Printed in Great Britain

Regressing Multivariate Gaussian Distribution on Vector Covariates for Co-expression Network Analysis

By ARNAB AUDDY

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.
Arnab.Auddy@pennmedicine.upenn.edu

T. TONY CAI

Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia Pennsylvania 19104, U.S.A.
tcai@wharton.upenn.edu

HONGZHE LI

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.
hongzhe@upenn.edu

Summary

Population-level single-cell gene expression data captures the gene expressions of thousands of cells for each individual within a sizable cohort. This data enables the construction of cell-type- and individual-specific gene co-expression network by estimating the covariance matrices. It is important to understand how such co-expression networks are associated with individual-level covariates. This paper considers Fréchet regression with multivariate Gaussian distribution as an outcome and vector covariates, where the Wasserstein distance between distributions is used as a replacement for the Euclidean distance. A test statistic is defined based on Fréchet mean and covariate weighted Fréchet mean. The asymptotic distribution of the test statistic is derived under the assumption of simultaneously diagonalizable covariance matrices. Although the proposed test statistic is motivated by considering the multivariate normal distribution as the outcome, it can be applied for testing the association between covariance matrices and covariates, where permutation can be used for assessing its statistical significance. Simulations show that the proposed test has correct type 1 error and adequate power. Results from an analysis of large-scale single-cell data reveal an association between the gene co-expression network of genes in the nutrient sensing pathway and age, indicating the perturbed gene co-expression network as people age.

Some key words: Fréchet mean, personalized co-expression, single cell gene expression, Wasserstein distance.

1. Introduction

The recent development of single-cell sequencing technologies, including both single cell RNA sequencing (scRNA-seq) and single nuclei sequencing, has made it possible to study cellular heterogeneity at high resolution and scale (Jovic et al., 2022; Tomokazu & Hafler, 2022). Cellular heterogeneity, which can be revealed by these single cell-level data, underlies phenotypic differences among individuals. Studying cellular heterogeneity is an important step toward our understanding of the disease onset, progression and treatment response. As the technologies mature and the cost is reduced, large-scale population level single cell data have been increasingly collected to characterize cellular heterogeneity from different aspects. These datasets enable the extraction of cell-type-specific personalized gene expression variations and molecular interactions, providing insights into changes in interaction networks with different covariates. However, linking such personalized single-cell data variability with disease phenotypes or covariates necessitates meticulous statistical analysis and the application of advanced methods.

Single-cell gene expression data enables the estimation of individual-specific, personalized gene co-expression networks within a set of genes (Ribeiro et al., 2022; Harris et al., 2021). The co-expression network among a set of genes can be approximated by a gene expression covariance matrix estimated based on the single cell gene expression data measured for a given individual. It is crucial to investigate how such co-expression networks are linked to individual covariates, such as age, sex, and genetic background.

This paper addresses the challenge of associating an individual-specific covariance matrix of a gene set with a vector of covariates within the framework of non-Euclidean outcome regression. For a given individual, we model the joint distribution of a gene set with a multivariate normal distribution, incorporating an individual-specific covariance matrix. Our goal is to perform regression analysis with multivariate normal distribution as the outcome for a set of Euclidean covariates. Such a regression model can the be used to test the association between gene expression covariance matrix and a set of covariates, and to address the question of how a covariate explains the variability of the covariance matrices observed over a large set of samples. Although the proposed test statistic is motivated by considering the multivariate normal distribution as the outcome, it can be applied for testing the association between covariance matrices and covariates, where permutation can be used for assessing its statistical significance.

Petersen & Müller (2019) considered regression relationships between responses that are complex random objects and vectors of real-valued predictors and developed Fréchet regression. They developed a global Fréchet regression relation as a generalization of multiple linear regression, as well as a class of more flexible local regression methods that generalizes local linear or polynomial regression. Their proposed regression approach for random objects incorporates the geometry implied by the metric and can be viewed as an extension of the Fréchet mean. Petersen et al. (2021) further developed Wasserstein F-test for association between a random distribution function and a covariate. For the special case of univariate density as the outcome, they derived the asymptotic null distribution, which can be used for testing the association between a random density function and a set of covariates. However, the methods and the theoretical results in Petersen & Müller (2019) and Petersen et al. (2021) assume that the random objects such as the univariate densities are observed without uncertainty.

In this paper, we consider Fréchet regression of Petersen & Müller (2019) for the setting where multivariate normal distribution is the outcome. In population scale single cell

studies, for each individual, we have an individual-specific gene expression distribution across different cells. Alongside, we have covariates (e.g., age, gender, disease status) that can possibly be connected to the joint distribution of gene expressions. Testing whether or not this connection is significant is important in practice since it allows us to determine the effect of covariates on the individual specific distributions. A primary reason of using the entire cell distribution, as obtained from single cell data, is its granularity. A pseudo-bulk single cell analysis, which reduces the distribution to a summary statistic (e.g., the sample mean) might not be sufficient to capture this dependence on covariates.

We note here that regression involving covariance matrices as responses have been previously considered in the literature, see, e.g., Hoff & Niu (2012); Zou et al. (2017). However these works consider a specific model which determines how the covariance matrix depends on the covariates. In this paper, we use a general framework based on Wasserstein distances between multivariate normal distributions which requires no such explicit model for the dependence of covariance matrices. The Wasserstein distance between two multivariate normal distributions, has an analytical expression (Givens & Shortt, 1984; Knott & Smith, 1984). Based on the results of Álvarez-Esteban et al. (2016); Chewi et al. (2020), the sample Wasserstein barycenter of multivariate normal distributions is also a multivariate normal with closed form expressions of its mean and covariance matrix. Although the test statistics we consider follow the general form of the Wasserstein F-test of Petersen et al. (2021), we show that, under simultaneously diagonalizable covariance matrices, the Wasserstein F-test follows a mixture of χ^2 limiting distribution under the null hypothesis of no association. We emphasize that our results do not assume that the multivariate normal densities are observed, instead, we allow these densities to be estimated using the observed data. Our results account for such an uncertainty. The assumption of simultaneously diagonalizable covariance matrices allow us to obtain more accurate estimate of the individual-specific covariance matrices. Additionally, we present the Wasserstein F-test for assessing the partial effect of a covariate while adjusting for other covariates.

2. Distributional Regression with Multivariate Normal Density Outcome

2.1. *Gene co-expression network from single cell data*

For typical single cell gene expression studies of n individuals, we observe the log-expression Y_{ijk} for the i^{th} individual, j^{th} cell and k^{th} gene. The indices run as $i = 1, \ldots, n$ $j = 1, \ldots, m_i$ and $k = 1, \ldots, d$. The model assumptions are as follows:

$$Y_{ij} \overset{iid}{\sim} N_d\left(\mu_i, \Sigma_i\right) \text{ for } j = 1, \ldots, m_i; i = 1, \ldots, n.$$

We have estimates

$$\widehat{\mu}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} Y_{ij}; \quad \widehat{\Sigma}_i = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(Y_{ij} - \widehat{\mu}_i)(Y_{ij} - \widehat{\mu}_i)^\top.$$

In genomics, Σ_i represents the individual-specific co-expression of d gene across the cells of a given type, and $\widehat{\Sigma}_i$ is its sample estimate. We also assume that we have p covariates measured, denoted by $X_i \in \mathbb{R}^p$ for the ith individual.

We will use the multivariate normal distribution as a motivating model for our work. This is different from the results in Petersen & Müller (2019) that are derived for univari-

ate distributions. In our applications to single cell genomics, it is necessary to consider multivariate distributions, and the Gaussian distribution is a natural starting point. While this motivates the specific form of the F-statistic we consider, we will find later that many of our results can in fact be generalized to other multivariate distributions provided some assumptions are satisfied.

2.2. *Fréchet regression with distribution as the outcome*

We begin with some preliminaries on regression where the responses are probability distributions. We will use the framework of Petersen & Müller (2019); Petersen et al. (2021), where the Wasserstein distance between two distributions is used as a replacement for the Euclidean distance used in typical regression scenarios with scalar responses.

Suppose that we observe response variables that are $N(\mu_i, \Sigma_i)$ distributions, together with covariates $X_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$. For two d-variate Gaussian distributions, say $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$, the Wasserstein distance between them can be shown to have the explicit form (Givens & Shortt, 1984; Knott & Smith, 1984):

$$d_W^2(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)) = \|\mu_1 - \mu_2\|_2^2 + \text{trace}(\Sigma_1 + \Sigma_2) - 2\text{trace}((\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}). \quad (1)$$

We follow the Fréchet regression approach of Petersen & Müller (2019); Petersen et al. (2021) as follows. Let us define the weights

$$s_{in}(x) = 1 + (X_i - \bar{X})^\top \hat{\Sigma}_X^{-1}(x - \bar{X}) \quad \text{for } i = 1, \ldots, n. \quad (2)$$

Given a set of weights we solve the weighted Wasserstein barycenter problem:

$$(\mu_*, \Sigma_*)(x) = \underset{\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d}_{\succeq \mathbf{0}}}{\text{argmin}} \sum_{i=1}^n s_{in}(x) d_W^2(N(\mu, \Sigma), N(\mu_i, \Sigma_i)).$$

By the results from Álvarez-Esteban et al. (2016); Chewi et al. (2020), the sample Wasserstein barycenter of $N(\mu_i, \Sigma_i)$ for $i = 1, \ldots, n$ is given by a normal distribution $N(\mu_0, \Sigma_0)$. Moreover, the parameters of the barycenter distribution satisfy:

$$\mu_0 = \frac{1}{n}\sum_{i=1}^n \mu_i \quad (3)$$

and Σ_0 satisfies the fixed-point equation

$$\Sigma_0 = \frac{1}{n}\sum_{i=1}^n (\Sigma_0^{1/2}\Sigma_i\Sigma_0^{1/2})^{1/2}. \quad (4)$$

In the context of regression, we consider a weighted barycenter problem, with weights $s_{in}(x)$ depending on the covariates x. Once again using the Gaussian Wasserstein barycenter results of Álvarez-Esteban et al. (2016); Chewi et al. (2020), we have

$$\mu_*(x) = \frac{1}{\sum_i s_{in}(x)} \sum_{i=1}^n s_{in}(x)\mu_i \quad (5)$$

while $\Sigma_*(x)$ satisfies the fixed-point equation

$$\Sigma_* = \frac{1}{\sum_i s_{in}(x)} \sum_{i=1}^n s_{in}(x)(\Sigma_*^{1/2}\Sigma_i\Sigma_*^{1/2})^{1/2}. \quad (6)$$

In (2), we obtain $s_{in}(\bar{X}) = 1$ and hence $(\mu_*, \Sigma_*)(\bar{X}) = (\mu_0, \Sigma_0)$, the sample mean Wasserstein barycenter given in equations (3) and (4).

3. Wasserstein F-test for no effects of covariates on multivariate normal distribution

3.1. *Wasserstein F-test*

We consider the problem of testing for the effect of covariates on the distribution valued response variables. More specifically, suppose for $i = 1, \ldots, n$ we observe F_i, which are distributions of random vectors $Y_i \in \mathbb{R}^d$. We consider the hypothesis of whether covariates X_i influence F_i. Motivated by our applications in gene co-expression network analysis, we will consider a parametric form of F_i, in particular, we will assume that $F_i \equiv N(\mu_i, \Sigma_i)$ for $i = 1, \ldots, n$. While this assumption might seem restrictive in practice, we use this only to motivate a specific form of the test statistic. In later sections, we show that, one can remove the normality assumption and still use the same test statistic to infer about certain forms of dependence of F_i on X_i.

We want to test the hypothesis that regressing on X_i has no effect on the multivariate distribution. Since a Gaussian distribution is completely specified by its mean and variance, this is equivalent to testing

$$H_0 : (\mu_*, \Sigma_*)(x) = (\mu_0, \Sigma_0) \text{ for all } x.$$

By equation (3.1) of Petersen et al. (2021), a natural test statistic is

$$F_* = \sum_{i=1}^{n} d_W^2(N((\mu_*, \Sigma_*)(X_i)), N((\mu_*, \Sigma_*)(\bar{X}))), \tag{7}$$

where $\mu_*(x)$ and $\Sigma_*(x)$ are as defined in equations (5) and (6). By the form of Wasserstein distance between Gaussian distributions (see equation (1)) we have

$$\begin{aligned} F_* = & \sum_{i=1}^{n} \left\| \mu_*(X_i) - \mu_*(\bar{X}) \right\|_2^2 + \sum_{j=1}^{n} \operatorname{trace}(\Sigma_*(X_j)) + n\operatorname{trace}(\Sigma_*(\bar{X})) \\ & - 2\sum_{j=1}^{n} \operatorname{trace}\left\{ ((\Sigma_*^{1/2}(X_j))(\Sigma_*(\bar{X}))(\Sigma_*^{1/2}(X_j)))^{1/2} \right\}. \end{aligned} \tag{8}$$

To better understand the effect of the covariates, let us define the weights

$$w_{ij} := \frac{s_{in}(X_j)}{\sum_{i=1}^{n} s_{in}(X_j)} = \frac{1 + (X_i - \bar{X})^\top \widehat{\Sigma}_X^{-1}(X_j - \bar{X})}{n},$$

and their centered versions

$$\check{w}_{ij} := \frac{s_{in}(X_j)}{\sum_{i=1}^{n} s_{in}(X_j)} - \frac{1}{n} = \frac{(X_i - \bar{X})^\top \widehat{\Sigma}_X^{-1}(X_j - \bar{X})}{n}. \tag{9}$$

We define the matrix of means

$$\mathbf{M}_\mu := (\mu_1\, \mu_2\, \ldots\, \mu_n) \in \mathbb{R}^{d \times n}. \tag{10}$$

Then the first term of F_* can then be written as

$$\sum_{i=1}^{n} \left\| \mu_*(X_i) - \mu_*(\bar{X}) \right\|_2^2 = \sum_{i=1}^{n} \check{w}_{\cdot i}^\top \mathbf{M}_\mu^\top \mathbf{M}_\mu \check{w}_{\cdot i} = \operatorname{trace}\left(\mathbf{M}_\mu^\top \mathbf{M}_\mu \sum_{i=1}^{n} \check{w}_{\cdot i} \check{w}_{\cdot i}^\top \right), \tag{11}$$

where $\check{w}_{\cdot i} = (\check{w}_{1i}, \ldots, \check{w}_{ni}) \in \mathbb{R}^n$ is the vector of centered weights at X_i.

3.2. *Simultaneously Diagonalizable Covariance Matrices*

We next consider the effect of covariates X_i on the covariances. Our results will be derived under a structural assumption that the covariance matrices $\Sigma_i \in \mathbb{R}^{d\times d}$ are simultaneously diagonalizable, for $i = 1, \ldots, n$. In particular, they share the same matrix of eigenvectors. Specifically,

$$\Sigma_i = U\Lambda_i U^\top \quad \text{for } i = 1, \ldots, n.$$

Here $\Lambda_i \in \mathbb{R}^{d\times d}$ are diagonal matrices of eigenvalues; and $U \in \mathbb{R}^{d\times d}$ is an orthonormal matrix of common eigenvectors. This assumption is also called common principal components analysis in the literature (see Flury (1984, 1986)). In our single cell application, this assumes that the underlying factors affecting the gene expressions are the same across different individuals, which is very plausible and is also verified in our data analysis in Section 6.

We first show that the solution $\Sigma_*(x)$ to equation (6) also has the same eigenmatrix, i.e., $\Sigma_*(x) = U\Lambda_*(x)U^\top$ for a diagonal $\Lambda_*(x) \in \mathbb{R}^{d\times d}$. If this holds, we then rewrite equation (6) to get

$$\begin{aligned} U\Lambda_* U^\top &= \frac{1}{n}\sum_{i=1}^n s_{in}(x)(\Sigma_*^{1/2}\Sigma_i\Sigma_*^{1/2})^{1/2} \\ &= \frac{1}{n}\sum_{i=1}^n s_{in}(x)(U\Lambda_*^{1/2}\Lambda_i\Lambda_*^{1/2}U^\top)^{1/2} \\ &= U\left(\frac{1}{n}\sum_{i=1}^n s_{in}(x)(\Lambda_*\Lambda_i)^{1/2}\right)U^\top, \end{aligned}$$

which implies

$$\Lambda_* = \frac{1}{n}\sum_{i=1}^n s_{in}(x)(\Lambda_*\Lambda_i)^{1/2} = \Lambda_*^{1/2}\left(\frac{1}{n}\sum_{i=1}^n s_{in}(x)\Lambda_i^{1/2}\right)$$

where the last line follows since Λ_* and Λ_i are diagonal matrices. We thus obtain the explicit form

$$\Lambda_*(x) = \left(\frac{1}{n}\sum_{i=1}^n s_{in}(x)\Lambda_i^{1/2}\right)^2$$

and

$$\Sigma_*(x) = U\left(\frac{1}{n}\sum_{i=1}^n s_{in}(x)\Lambda_i^{1/2}\right)^2 U^\top. \tag{12}$$

This implies, to satisfy the fixed point equation, $\Sigma_*(x)$ indeed has the same eigenmatrix U.

We denote the vectors of eigenvalues of $\Sigma_i^{1/2}$ by

$$\gamma_i := \left(\sqrt{\lambda_{i1}}\ \ldots, \sqrt{\lambda_{id}}\right) = \mathrm{diag}(U^\top\Sigma_i^{1/2}U).$$

and the matrix

$$\Gamma = (\gamma_1\ \gamma_2\ \ldots\ \gamma_n) \in \mathbb{R}^{d\times n}. \tag{13}$$

We also define the matrix of scaled and centered covariates:

$$Z_X = \left(\widehat{\Sigma}_X^{-1/2}(X_1 - \bar{X}) \ \ldots \ \widehat{\Sigma}_X^{-1/2}(X_n - \bar{X})\right) \in \mathbb{R}^{p\times n}. \tag{14}$$

We then have the following theorem that gives an explicit expression for the F-statistic.

THEOREM 1. *For $i = 1, \ldots, n$ if Σ_i are simultaneously diagonalizable, F_* defined in equation (8) can be written as $F_* = \frac{1}{n}\|M_\mu Z_X^\top\|_{\mathrm{F}}^2 + \frac{1}{n}\|\Gamma Z_X^\top\|_{\mathrm{F}}^2$, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix.*

3.3. *Estimating simultaneously diagonalizable covariance matrices*

Computation of the F-statistic derived in the previous section requires the knowledge of μ_i and Σ_i. In practice we only observe samples from the $N(\mu_i, \Sigma_i)$ distributions, and have to estimate μ_i and Σ_i from the data. To this end, suppose we have i.i.d. observations $(Y_{i1}, \ldots, Y_{im_i})$ for $i = 1, \ldots, n$. We now describe the estimation of μ and Σ under the additional structural assumption imposed on the covariance matrices.

Under the assumption of diagonalizable covariance matrices, $\Sigma_i = U\Lambda_i U^\top$, where $U \in \mathbb{R}^{d\times d}$ is the matrix of common eigenvectors. It follows that

$$\Sigma_* = \frac{1}{n}\sum_{i=1}^n \Sigma_i = U\left(\frac{1}{n}\sum_{i=1}^n \Lambda_i\right)U^\top.$$

To leverage this additional structure of shared eigenvectors, we therefore estimate Σ_i in the following manner:

1. Split the samples into two halves and denote them by $Y_{ij}^{(1)}$ and $Y_{ij}^{(2)}$.
2. Estimate Σ_* by pooling all the samples in the first half, i.e.,

$$\widehat{\Sigma}_* = \frac{1}{\sum m_i}\sum_{i=1}^n\sum_{j=1}^{m_i}\left(Y_{ij}^{(1)} - \bar{\bar{Y}}\right)\left(Y_{ij}^{(1)} - \bar{\bar{Y}}\right)^\top. \tag{15}$$

3. Estimate U by $\widehat{U}_*$, which are the left singular vectors of $\widehat{\Sigma}_*$.
4. Estimate the vector of eigenvalues of Σ_i via:

$$\widehat{\Lambda}_i := \mathrm{diag}\left(\left\{\widehat{\lambda}_{ik} : 1 \le k \le d\right\}\right) \ \text{ where } \ \widehat{\lambda}_{ik} := \frac{1}{m_i}\sum_{j=1}^{m_i}(\widehat{\mathbf{u}}_k^\top \mathbf{Y}_{ij}^{(2)})^2 \ \text{ for } \ 1 \le k \le d. \tag{16}$$

5. Estimate Σ_i as follows:

$$\widehat{\Sigma}_i = \widehat{U}_*^\top \widehat{\Lambda}_i \widehat{U}_*. \tag{17}$$

Now in the definition of F_*, we replace all occurrences of the true, unknown Σ_i by the $\widehat{\Sigma}_i$ and its related quantities defined above. In particular, let us define

$$\widehat{F}_* = \sum_{i=1}^n d_W^2(N((\widehat{\mu}_*, \widehat{\Sigma}_*)(X_i)), N((\widehat{\mu}_*, \widehat{\Sigma}_*)(\bar{X}))). \tag{18}$$

For simplicity let us assume that Y_{ij} are suitably centered so that

$$\mu_1 = \mu_2 = \cdots = \mu_n = 0.$$

Note that we can estimate Γ in (13) by

$$\widehat{\Gamma} = (\widehat{\gamma}_1 \ \widehat{\gamma}_2 \ \ldots \ \widehat{\gamma}_n) \in \mathbb{R}^{d\times n} \tag{19}$$

where

$$\widehat{\gamma}_i = \text{vec}(\text{diag}(\widehat{\Lambda}_i)^{\frac{1}{2}}) \in \mathbb{R}^d \tag{20}$$

for $i = 1, \ldots, n$. Moreover we will write

$$\widehat{\gamma}_{k\cdot} \in \mathbb{R}^n \text{ for } k = 1, \ldots, d$$

to denote the rows of $\widehat{\Gamma}$. Then following the calculations leading to the alternative definition of F_* in Theorem 1, we have the following proposition.

PROPOSITION 1. *For $i = 1, \ldots, n$ if Σ_i are simultaneously diagonalizable, $\widehat{F}_*$ defined in equation (18) can be written as*

$$\widehat{F}_* = \sum_{s=1}^{p} \sum_{k=1}^{d} \left(\sum_{i=1}^{n} (v_{X,s})_i \sqrt{\widehat{\lambda}_{ik}} \right)^2,$$

where $v_{X,s}$ are the eigenvectors of $Z_X^\top Z_X$.

We are now ready to present the main result of this section, namely, the null distribution of $\widehat{F}_*$. We remind the reader that our null hypothesis is $H_0 : \Sigma(x) = \Sigma_0$ for all covariate values $x \in \mathbb{R}^p$. Let Σ have the eigendecomposition

$$H_0 : \Sigma(x) = \Sigma_0 \text{ for all } x \in \mathbb{R}^p; \Sigma_0 = U \Lambda U^\top.$$

The following theorem presents the null distribution of $\widehat{F}_*$.

THEOREM 2. *Under H_0 as described above, and $\widehat{F}_*$ defined in equation (18), there exists a constant $C > 0$ such that*

$$\mathbb{P}\left(\widehat{F}_* \leq x\right) \in \left[\mathbb{P}\left\{ \tilde{F}_* \leq x \left(1 - C\sqrt{\frac{d}{\sum_i m_i}}\right) \right\}, \mathbb{P}\left\{ \tilde{F}_* \leq x \left(1 + C\sqrt{\frac{d}{\sum_i m_i}}\right) \right\} \right] \tag{21}$$

for all $x \in \mathbb{R}$. Here $\tilde{F}_* = \sum_{k=1}^{d} \sum_{s=1}^{p} \eta_{ks}^* (\chi_1^2)_{ks}$, where $(\chi_1^2)_{ks}$ for $k = 1, \ldots, d$ and $s = 1, \ldots, p$ denote $d \times p$ i.i.d. χ_1^2 random variables. Moreover, η_{ks}^* are the eigenvalues of $Z_X D_k^* Z_X^\top$, where

$$D_k^* = \text{diag}\left(\left\{ \frac{\lambda_k}{2m_1} \ldots \frac{\lambda_k}{2m_n} \right\} \right). \tag{22}$$

Finally, λ_k are the eigenvalues of Σ_0, i.e, the common covariance matrix under H_0.

We add a few remarks here to interpret the results of the previous theorem.

Remark 1. (Asymptotic null distribution) When $\sum_i m_i \gg d$, it follows from the above theorem that

$$\widehat{F}_* \xrightarrow{\mathcal{L}} \sum_{k=1}^{d} \sum_{s=1}^{p} \eta_{ks}^* (\chi_1^2)_{ks},$$

where $(\chi_1^2)_{ks}$ for $k = 1, \ldots, d$ and $s = 1, \ldots, p$ denote dp many i.i.d. χ_1^2 random variables.

Remark 2. (Dependence on sample size) Theorem 2 provides a non-asymptotic version of the convergence of $\widehat{F}_*$, in terms of the sample sizes $\sum_i m_i$ and the dimension d. We recognize an advantage of using the simultaneously diagonalizable assumption here. If

we were to use the individual covariance matrices to consistently estimate their population counterparts, the distribution convergence would require $m_i \gg d$ for all $1 \leq i \leq n$. However, under our assumption, it suffices to estimate the common eigenvectors from the pooled covariance matrix, which is possible as soon as $\sum_i m_i \gg d$.

Remark 3. (Single covariate) A case of special interest is $p = 1$. In this case, the null distribution has the simpler expression $\sum_{k=1}^{d} \eta_k^* (\chi_1^2)_k$ in terms of the eigenvalues

$$\eta_k^* = \frac{\lambda_k}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{2m_i} \tag{23}$$

for $k = 1, \ldots, d$.

Remark 4. (Equal sample sizes) Suppose that the individual specific sample sizes are all equal, i.e., $m_1 = m_2 = \cdots = m_n = m$. In this case the null distribution simplifies to $\sum_{k=1}^{d} \eta_k^* (\chi_p^2)_k$ in terms of the eigenvalues

$$\eta_k^* = \frac{n\lambda_k}{2m} \tag{24}$$

for $k = 1, \ldots, d$.

Although our test statistic has been derived as the Wasserstein distance between two Normal distributions, we can treat F_*, as defined in Theorem 1, as a statistic for testing the effect of covariates on the mean and covariance matrix of a random vector $Y \in \mathbb{R}^d$. In particular, the normality assumption is not required if we intend to use F_* in this manner. More interestingly, the asymptotic null distribution of $\widehat{F}_*$ holds true more generally, even when the normality assumption is violated. It remains the same as long as the common covariance matrix is Σ_0 under H_0, and the eigenvectors of $(D_k^*)^{1/2} Z_X^\top Z_X (D_k^*)^{1/2}$, for $k = 1, \ldots, d$, are sufficiently incoherent. Here $D_k^* \in \mathbb{R}^{d \times d}$ are as defined in (22).

Thus, although the assumption that observations are normal is too restrictive in practice, we can bypass this issue and use the equivalent form of F_* and $\widehat{F}_*$ and still determine the effect of covariates. It is important to keep in mind that in this new form, we are only able to capture the influence of the predictors on the mean and variance of Y, as opposed to the entire distribution of Y, for which one requires the full strength of the Wasserstein distance based Fréchet regression introduced in Petersen & Müller (2019); Petersen et al. (2021).

4. Wasserstein F-test for Partial Effects of Covariates

Using the same setup, we can also test for presence of partial effects of regression. In particular, suppose $x \in \mathbb{R}^p$ can be split into two sub-vectors $(x^{(1)}, x^{(2)})$. Here

$$x^{(1)} \in \mathbb{R}^{p_1} \text{ and } x^{(2)} \in \mathbb{R}^{p_2} \text{ with } p = p_1 + p_2.$$

We want to test the hypothesis that $x^{(2)}$ has no effect, i.e.,

$$H_0^P : (\mu_*, \Sigma_*)(x) = (\mu_*, \Sigma_*)(x^{(1)})$$

versus $H_1 : H_0^P$ is not true. Following Section 3.2 of Petersen et al. (2021), a suitable statistic for testing this hypothesis is given by

$$F_*^P = \sum_{i=1}^{n} d_W^2(N((\mu_*, \Sigma_*)(X_i)), N((\mu_*, \Sigma_*)(\bar{X}))) - \sum_{i=1}^{n} d_W^2(N((\mu_*, \Sigma_*)(X_i^{(1)})), N((\mu_*, \Sigma_*)(\bar{X}^{(1)}))). \tag{25}$$

Following (14), we define

$$Z_{X^{(1)}} = \left(\widehat{\Sigma}_{X^{(1)}}^{-1/2}(X_1^{(1)} - \bar{X}^{(1)}) \ \ldots \ \widehat{\Sigma}_{X^{(1)}}^{-1/2}(X_n^{(1)} - \bar{X}^{(1)})\right) \in \mathbb{R}^{p_1 \times n}. \tag{26}$$

For simplicity let us assume that Y_{ij} are suitably centered so that $\mu_1 = \mu_2 = \cdots = \mu_n = 0$. Then following the computation behind (32), we get

$$\widehat{F}_*^P = \frac{1}{n}\left\|\widehat{\Gamma} Z_X^\top\right\|_{\mathrm{F}}^2 - \frac{1}{n}\left\|\widehat{\Gamma} Z_{X^{(1)}}^\top\right\|_{\mathrm{F}}^2 = \frac{1}{n}\sum_{k=1}^{d} \widehat{\gamma}_{k\cdot}^\top \left(Z_X^\top Z_X - Z_{X^{(1)}}^\top Z_{X^{(1)}}\right)\widehat{\gamma}_{k\cdot}. \tag{27}$$

The following proposition furnishes an alternative expression for $\widehat{F}_*^P$, which will be more amenable to our existing framework for deriving null distributions.

PROPOSITION 2. *For $i = 1, \ldots, n$ if Σ_i are simultaneously diagonalizable, $\widehat{F}_*^P$ defined in equation (27) can be written as*

$$\widehat{F}_*^P = \frac{1}{n}\sum_{k=1}^{d} \widehat{\gamma}_{k\cdot}^\top Z_{X^{(2|1)}}^\top Z_{X^{(2|1)}} \widehat{\gamma}_{k\cdot}.$$

where $Z_{X^{(2|1)}} \in \mathbb{R}^{p_2 \times n}$ is a matrix whose s-th column is equal to

$$(Z_{X^{(2|1)}})_s := \Sigma_{22|1}^{-1/2}\left(\Sigma_{12}^\top \Sigma_{X^{(1)}}^{-1}(X_s^{(1)} - \bar{X}^{(1)}) - (X_s^{(2)} - \bar{X}^{(2)})\right) \in \mathbb{R}^{p_2}, \ \text{for } s = 1, \ldots, n.$$

Then as before we use the eigendecomposition of $Z_X^\top Z_X - Z_{X^{(1)}}^\top Z_{X^{(1)}}$ to conclude that the null distribution will be a mixture of $d \times p_2$ i.i.d. χ_1^2 random variables. In particular, we have the following theorem.

THEOREM 3. *Under H_0^P as described above, and $\widehat{F}_*^P$ defined in equation (25), Under H_0 as described above, and $\widehat{F}_*$ defined in equation (18), there exists a constant $C > 0$ such that*

$$\mathbb{P}\left(\widehat{F}_* \le x\right) \in \left[\mathbb{P}\left\{\tilde{F}_*^P \le x\left(1 - C\sqrt{\frac{d}{\sum_i m_i}}\right)\right\}, \mathbb{P}\left\{\tilde{F}_*^P \le x\left(1 + C\sqrt{\frac{d}{\sum_i m_i}}\right)\right\}\right] \tag{28}$$

for all $x \in \mathbb{R}$. Here $\tilde{F}_*^P = \sum_{k=1}^{d}\sum_{s=1}^{p_2} \eta_{ks,(2)|(1)}^* (\chi_1^2)_{ks}$, where $(\chi_1^2)_{ks}$ for $k = 1, \ldots, d$ and $s = 1, \ldots, p_2$ denote dp_2 many i.i.d. χ_1^2 random variables, and $\eta_{ks,(2)|(1)}^*$ are the eigenvalues of $Z_{X^{(2)|(1)}} D_k^* Z_{X^{(2)|(1)}}^\top$, where

$$D_k^* = \operatorname{diag}\left(\left\{\frac{\lambda_{1k}}{2m_1} \ldots \frac{\lambda_{nk}}{2m_n}\right\}\right).$$

Finally, λ_{ik} are the k-th eigenvalues of Σ_i, for $k = 1, \ldots, d$, where Σ_i, for $i = 1, \ldots, n$ are the covariance matrices under H_0^P.

5. Numerical Experiments

The proposed test is computationally is easy to implement. In this section, we investigate its numerical performance and practical implications in a simulation study.

In the first set of simulation experiments, we test the accuracy of the null distribution of the F-statistic for no effects. The simulation setting was as follows. We fix $d = 50$, $p = 2$, $n = 800$ and $m_i = 200$ for $i = 1, \ldots, n$. In the first simulation setting, we use the null model where there is no effect of covariates. The true eigenvalues of all the covariance matrices Σ_i are taken to be 5 and 2 with multiplicity 25 each. The common matrix of eigenvectors is simulated from the Haar measure on orthonormal matrices. The QQ-plot of $\widehat{F}_*$ against the theoretically derived null distribution is presented in Figure 1 (a), and shows a very close match of the observed and theoretically obtained quantiles.

We then assess the power of the F-test in the second simulation setting. We again use the same values of d, p, n and m_i. The common matrix of eigenvectors is simulated from the Haar measure on orthonormal matrices. The true eigenvalues of all the covariance matrices Σ_i are taken to be λ_1 and λ_2 with multiplicity 25 each. Here $\lambda_1 = 5 + \delta X_1 + \delta X_2$ and $\lambda_2 = 2 + \delta X_1 + \delta X_2$, where $X_1, X_2 \sim \text{Uniform}(21, 70)$ are the covariates. The power of the test as a function of δ is plotted in Figure 1 (b).

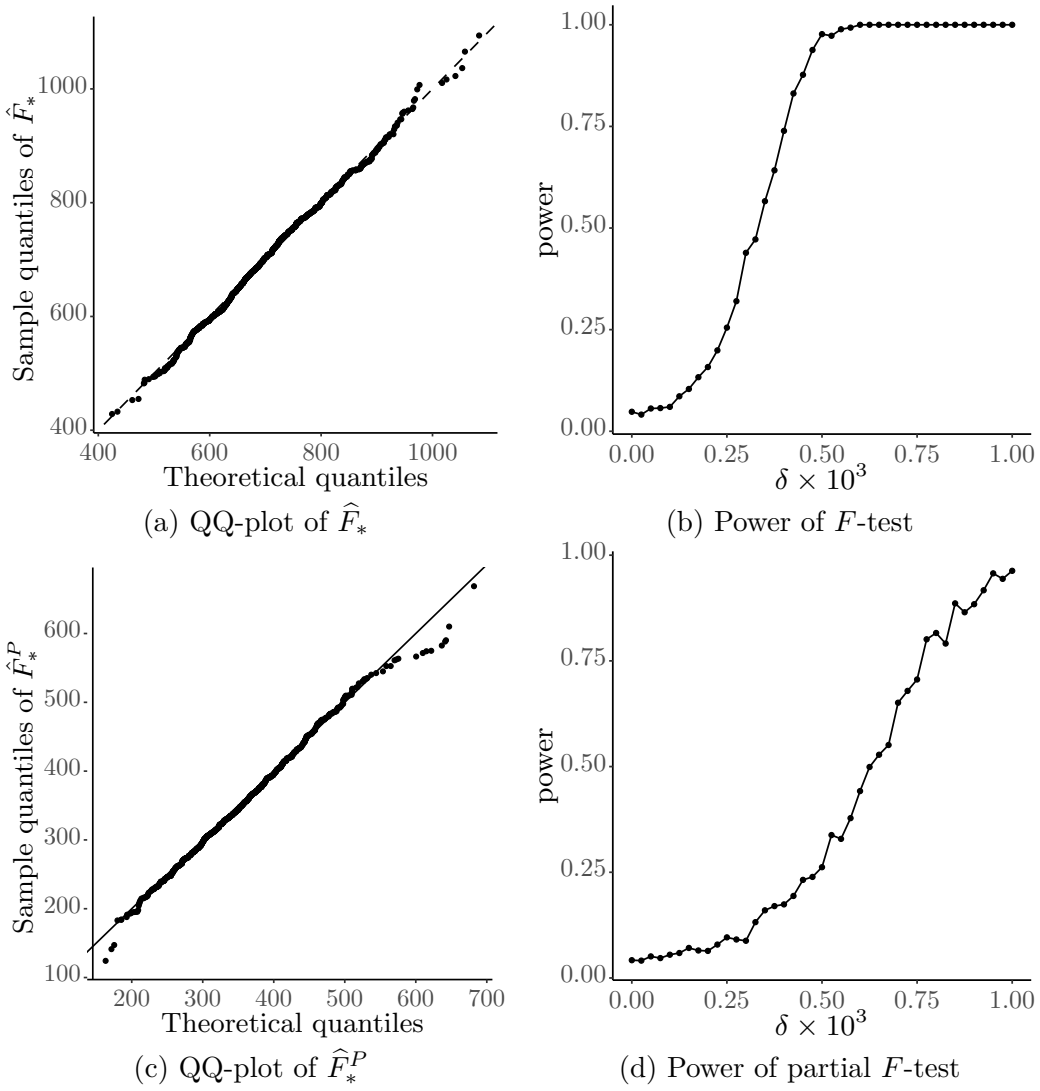
In the second set of simulation experiments, we test the accuracy of the null distribution of the F-statistic for partial effects. The simulation setting was exactly the same as in the case of testing for no effects. In the first experiment, we use the partial null model where there is only an effect of the first covariate. The true eigenvalues of all the covariance matrices Σ_i are taken to be 0.001 times X_1. The common matrix of eigenvectors is simulated from the Haar measure on orthonormal matrices. The QQ-plot of $\widehat{F}_*$ against the theoretically derived null distribution is presented in Figure 1 (c) , and shows a very close match of the observed and theoretically obtained quantiles.

We assess the power of the partial F-test in the second simulation setting. We again use the same values of d, p, n and m_i. The common matrix of eigenvectors is simulated from the Haar measure on orthonormal matrices. The true eigenvalues of all the covariance matrices Σ_i are taken to be λ_1 and λ_2 with multiplicity 25 each. Here $\lambda_1 = 5 + 0.001X_1 + \delta X_2$ and $\lambda_2 = 2 + 0.001X_1 + \delta X_2$, where $X_1, X_2 \sim \text{Uniform}(21, 70)$ are the covariates. The power of the partial F-test as a function of δ is plotted in Figure 1 (d).

6. Analysis of single cell gene expressions across different ages

Aging is a complex process of accumulation of molecular, cellular, and organ damage, leading to loss of function and increased vulnerability to disease and death. Nutrient-sensing pathways, namely insulin/insulin-like growth factor signaling and target-of- rapamycin can substantially increase healthy life span of laboratory model organisms (Davinelli et al., 2012; de Lucia et al., 2020). These nutrient signaling pathways are conserved in various organisms. We are interested in understanding the co-expression structure of 61 genes in this KEGG nutrient-sensing pathways based on the recently published population scale single cell RNA-seq data of human peripheral blood mononuclear cells (PBMCs) from blood samples of over 982 healthy individuals with ages ranging from 20 to 90 (Yazar et al., 2022).

We focus our analysis on CD4+ naive and central memory T (CD4NC) cells, which is the most common cell type observed in the data. Age-associated changes in CD4 T-cell functionality have been linked to chronic inflammation and decreased immunity (Elyahu

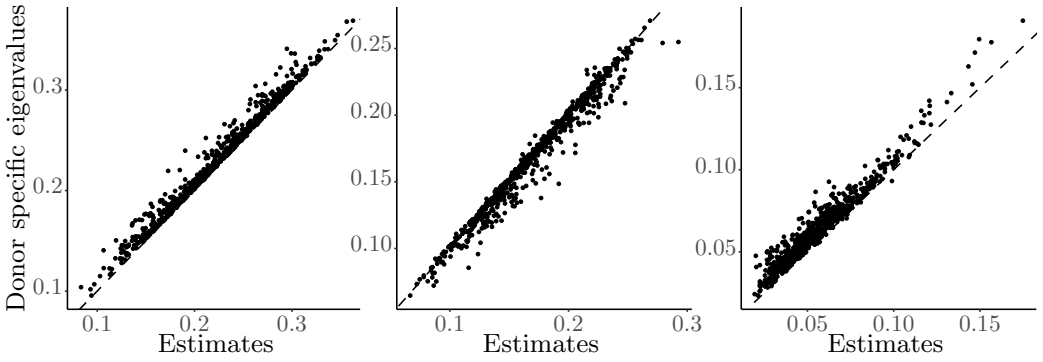


(a) QQ-plot of $\widehat{F}_*$ (b) Power of F-test

(c) QQ-plot of $\widehat{F}_*^P$ (d) Power of partial F-test

Fig. 1: (a) and (b): Performance of the F-test when testing for no effects on simulated data; (c) and(d): Performance of the F-test when testing for partial effects on simulated data.

et al., 2019). There are a total of 50 genes that are expressed in this cell type. We use the F-test in Section 3 to test whether the gene distributions across different cells changes with age of an individual. We select 734 donors for each of whom at least 200 cells are observed. For each individual and each cell, we have the gene expression across 50 genes, so that they are expressed in at least one cell for every individual. Alongside we have the age of each individual. The gene expression data is very sparse. To reduce the dimensionality of the data, we further select the 26 genes which have been expressed in at least 3% of the cells, across different donors. We note here that the gene expression data is extremely sparse, and we use this fact when centering the data. That is, we center the nonzero values around the mean of the nonzero values, while the zero values are left



(a) First eigenvalue (b) Second eigenvalue (c) Third eigenvalue

Fig. 2: Plot of the first three eigenvalues of donor-specific covariance matrices: the Y-axis has the actual eigenvalues computed directly from covariance matrices, while the X-axis has the estimates obtained from (16) under the simultaneous diagonalizability assumption.

unchanged. Then the individual as well as the grand covariance matrices are calculated with respect to this centered dataset.

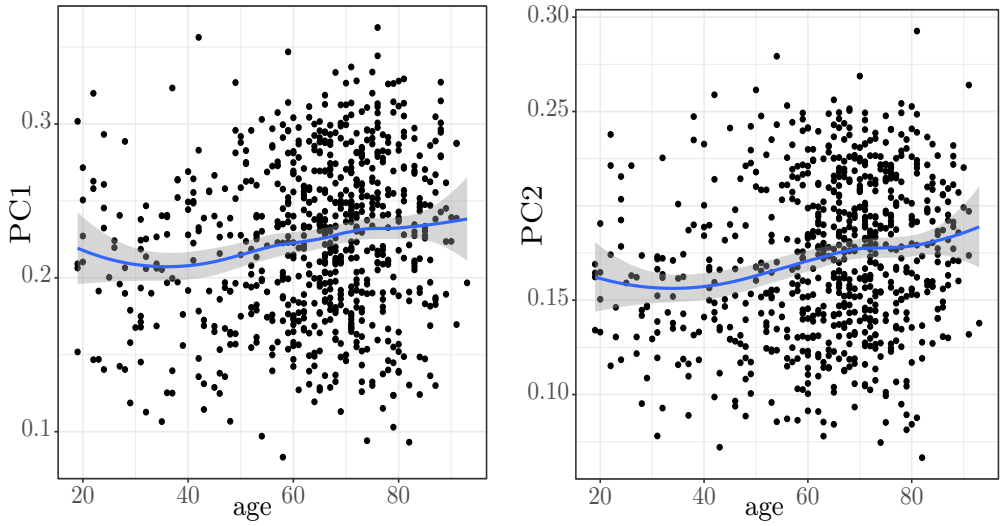
Before presenting our results, we provide some empirical justification behind using the assumption of simultaneous diagonalizability of the covariance matrices. Figure 2 shows the top three eigenvalues of the individual specific covariance matrices with and without the assumption of the simultaneous diagonalizability. That is, each point represents $(\widehat{\lambda}_{ik}, \tilde{\lambda}_{ik})$ for $k = 1, 2, 3$ and $i = 1, \ldots, 734$. The X-axis plots the estimates $\widehat{\lambda}_{ik}$, i.e., the k-th eigenvalue of $\widehat{\Sigma}_i$ under the assumption of simultaneous diagonalizability, as calculated in (16). On the other hand, the Y-axis plots the values $\tilde{\lambda}_{ik}$, which are the k-th eigenvalue calculated directly from $\widehat{\Sigma}_i$'s. The line $y = x$ is added for reference. Each subplot in Figure 2 shows that the estimated eigenvalues (under the common eigenvector assumption) are reasonably close to the true eigenvalues.

Next, we perform the F-test of no effects to test whether the covariance matrices have a significant dependence on age. With the test statistic computed in Section 3, we perform the test of no effects with i) the null distribution derived in Theorem 2 as well as ii) permutation based test. Both tests of no effects are rejected with a p-value of 0. The value of the F-statistic is 199.04. The permutation test is performed with 10000 random permutations.

In Figure 3 we separately plot the first two eigenvalues against age together with the Lowess fits of the data. It is evident that for higher age, the eigenvalues take larger values, indicating overall changes of covariances among these genes. In addition, we observe a larger variability of the first and second eigenvalues for the older individuals. This implies larger variability of co-expressions in older ages than younger ages.

7. Discussion

This paper has developed test for association between multivariate normal distribution and a set of covariates in the framework of Fréchet regression. We specifically considered the setting that the covariance matrices are simultaneously diagonalizable, which enable



(a) First eigenvalue (b) Second eigenvalue

Fig. 3: The above figure plots the first two eigenvalues of donor-specific covariance matrices under the common eigenvector assumption. Both figures show a general increasing trend as age increases.

efficient estimation of all the covariance matrices. We have developed the Wasserstein F-tests for association between a multivariate normal distribution outcome and a set of covariates, which has a mixture of χ^2 distributions under the null.

Although the tests are developed treating the multivariate normal distribution as the outcome, the test statistic itself can still be applied without the normality assumption, in which case the covariance matrix itself is treated as an outcome. Under some technical assumptions, the limiting null distribution still holds. In many applications when the multivariate normality assumptions do not hold, one can apply certain transformations to ensure the normality.

8. Appendix - Proofs

Proof of Theorem 1. We analyze the last term in F_* next:

$$\begin{aligned}
&\operatorname{trace}\left\{((\Sigma_*^{1/2}(X_j))(\Sigma_*(\bar{X}))(\Sigma_*^{1/2}(X_j)))^{1/2}\right\}\\
&= \operatorname{trace}\left\{(\Lambda_*^{1/2}(X_j))(\Lambda_*(\bar{X}))(\Lambda_*^{1/2}(X_j))^{1/2}\right\}\\
&= \sum_{k=1}^{d}(\lambda_k(X_j)\lambda_k(\bar{X}))^{1/2}.
\end{aligned}$$

Consequently

$$\begin{aligned}
&\operatorname{trace}(\Sigma_*(X_j)) + \operatorname{trace}(\Sigma_*(\bar{X})) - 2\operatorname{trace}\left\{((\Sigma_*^{1/2}(X_j))(\Sigma_*(\bar{X}))(\Sigma_*^{1/2}(X_j)))^{1/2}\right\}\\
&= \sum_{k=1}^{d}(\sqrt{\lambda}_k(X_j) - \sqrt{\lambda}_k(\bar{X}))^2\\
&= \sum_{k=1}^{d}\left(\frac{1}{n}\sum_{i=1}^{n}(s_{in}(X_j) - s_{in}(\bar{X}))\lambda_{ik}^{1/2}\right)^2 = \sum_{k=1}^{d}\left(\sum_{i=1}^{n}\check{w}_{ij}\lambda_{ik}^{1/2}\right)^2 \qquad (29)
\end{aligned}$$

Now summing (29) over $j = 1, \ldots, n$ we have

$$\begin{aligned}
&\sum_{j=1}^{n}\left[\operatorname{trace}(\Sigma_*(X_j)) + \operatorname{trace}(\Sigma_*(\bar{X}))\right] - 2\left[\operatorname{trace}\left\{((\Sigma_*^{1/2}(X_j))(\Sigma_*(\bar{X}))(\Sigma_*^{1/2}(X_j)))^{1/2}\right\}\right]\\
&= \sum_{j=1}^{n}\sum_{k=1}^{d}\left(\sum_{i=1}^{n}\check{w}_{ij}\gamma_{ik}\right)^2 = \sum_{j=1}^{n}\check{w}_{\cdot j}^{\top}\mathbf{\Gamma}^{\top}\mathbf{\Gamma}\check{w}_{\cdot j} = \operatorname{trace}\left(\mathbf{\Gamma}^{\top}\mathbf{\Gamma}\sum_{i=1}^{n}\check{w}_{\cdot i}\check{w}_{\cdot i}^{\top}\right). \qquad (30)
\end{aligned}$$

We next compute the (j,k) entries

$$\begin{aligned}
\left(\sum_{i=1}^{n}\check{w}_{\cdot i}\check{w}_{\cdot i}^{\top}\right)_{jk} &= \frac{1}{n^2}(X_j - \bar{X})^{\top}\widehat{\Sigma}_X^{-1}\left(\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^{\top}\right)\widehat{\Sigma}_X^{-1}(X_k - \bar{X})\\
&= \frac{1}{n}(X_j - \bar{X})^{\top}\widehat{\Sigma}_X^{-1}(X_k - \bar{X})
\end{aligned}$$

for $1 \le j, k \le n$. In particular, we have the Gram matrix representation

$$\sum_{i=1}^{n}\check{w}_{\cdot i}\check{w}_{\cdot i}^{\top} = \frac{1}{n}Z_X^{\top}Z_X. \qquad (31)$$

By the definition of F_* from equation (8), and using equations (11), (29), and (30) we have

$$\begin{aligned}
&F_*\\
&= \operatorname{trace}\left(\left\{M_\mu^{\top}M_\mu + \Gamma^{\top}\Gamma\right\}\sum_{i=1}^{n}\check{w}_{\cdot i}\check{w}_{\cdot i}^{\top}\right)\\
&= \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n}(\langle\mu_j,\mu_k\rangle + \langle\gamma_j,\gamma_k\rangle)(X_j - \bar{X})^{\top}\widehat{\Sigma}_X^{-1}(X_k - \bar{X})\\
&= \frac{1}{n}\|M_\mu Z_X^{\top}\|_{\mathrm{F}}^2 + \frac{1}{n}\|\Gamma Z_X^{\top}\|_{\mathrm{F}}^2 \qquad (32)
\end{aligned}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix. □

Proof of Proposition 1. We have using (18) that

$$\begin{aligned}
\widehat{F}_* &= \frac{1}{n}\left\|\widehat{\Gamma} Z_X^\top\right\|_{\mathrm{F}}^2 = \frac{1}{n}\left\|Z_X \widehat{\Gamma}^\top\right\|_{\mathrm{F}}^2 \\
&= \frac{1}{n}\sum_{k=1}^{d} \|Z_X \widehat{\gamma}_{k\cdot}\|^2 = \frac{1}{n}\sum_{k=1}^{d} \widehat{\gamma}_{k\cdot}^\top Z_X^\top Z_X \widehat{\gamma}_{k\cdot} \\
&= \frac{1}{n}\sum_{s=1}^{p} \lambda_{X,s} \sum_{k=1}^{d} \left(v_{X,s}^\top \widehat{\gamma}_{k\cdot}\right)^2 \\
&= \sum_{s=1}^{p} \lambda_{X,s} \sum_{k=1}^{d} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (v_{X,s})_i \sqrt{\widehat{\lambda}_{ik}}\right)^2
\end{aligned} \tag{33}$$

where we use the eigendecomposition

$$Z_X^\top Z_X = \sum_{s=1}^{p} \lambda_{X,s} v_{X,s} v_{X,s}^\top.$$

We can determine the eigenvalues $\lambda_{X,s}$ as follows:

$$Z_X Z_X^\top = \widehat{\Sigma}_X^{-1/2} \left(\sum_{j=1}^{n} (X_j - \bar{X})(X_j - \bar{X})^\top\right) \widehat{\Sigma}_X^{-1/2} = n\mathbb{I}_p.$$

Since $Z_X Z_X^\top$ and $Z_X^\top Z_X$ have the same eigenvalues, it follows that

$$\lambda_{X,s} = n \quad \text{for } s = 1, \ldots, p.$$

Plugging in the last equality into equation (33) finishes the proof. □

Proof of Theorem 2. Note that by the definition in (16) it is clear that, conditional on $Y_{ij}^{(1)}$, $\{\widehat{\lambda}_{ik} : 1 \le i \le n; 1 \le k \le d\}$ are independent random variables, owing to the orthogonality of $\widehat{u}_k$ and the independence of Y_{ij}'s. Note that since the columns of Z_X have mean zero, we have $\sum_{i=1}^{n} (v_{X,s})_i = 0$.

Under H_0, we have

$$\mathbb{E}(\sqrt{\widehat{\lambda}}_{1k} | Y_{ij}^{(1)}) = \cdots = \mathbb{E}(\sqrt{\widehat{\lambda}}_{nk} | Y_{ij}^{(1)}).$$

and hence

$$\mathbb{E}\left(\sum_{i=1}^{n} (v_{X,s})_i \sqrt{\widehat{\lambda}}_{ik} | Y_{ij}^{(1)}\right) = 0. \tag{34}$$

Moreover, by delta method,

$$\begin{aligned}\sigma_{ik}^2 := \mathrm{Var}\left(\sqrt{\widehat{\lambda}}_{ik}|Y_{ij}^{(1)}\right) &= \mathrm{Var}\left(\sqrt{\frac{1}{m_i}\sum_{j=1}^{m_i}(\widehat{u}_k^\top Y_{ij}^{(2)})^2}|Y_{ij}^{(1)}\right) \\ &= \frac{1}{4m_i\widehat{u}_k^\top\Sigma_0\widehat{u}_k}\mathrm{Var}\left((\widehat{u}_k^\top Y_{i1}^{(2)})^2|Y_{ij}^{(1)}\right) + O(m_i^{-2}) \\ &= \frac{\widehat{u}_k^\top\Sigma_0\widehat{u}_k}{2m_i} + O(m_i^{-2}).\end{aligned}$$

Thus, under H_0 we have

$$\widehat{F}_* = \sum_{k=1}^{d}\widehat{\gamma}_{k\cdot}^\top Z_X^\top Z_X\widehat{\gamma}_{k\cdot} = \sum_{k=1}^{d}\left(\widehat{\gamma}_{k\cdot} - \mathbb{E}\widehat{\gamma}_{k\cdot}\right)^\top Z_X^\top Z_X\left(\widehat{\gamma}_{k\cdot} - \mathbb{E}\widehat{\gamma}_{k\cdot}\right)$$

where the second equality follows from equation (34). Defining $D_k := \mathrm{diag}(\sigma_{1k}^2\,\sigma_{2k}^2\,\ldots\,\sigma_{nk}^2)$, it follows from the independence of the random variables $\{\widehat{\lambda}_{ik} : 1 \le i \le n; 1 \le k \le d\}$ that under H_0

$$(\widehat{\gamma}_{k\cdot} - \mathbb{E}\widehat{\gamma}_{k\cdot}) \sim N(0, D_k), \text{ for } 1 \le k \le d \tag{35}$$

conditional on $Y_{ij}^{(1)}$. We can then write

$$\begin{aligned}&\left(\widehat{\gamma}_{k\cdot} - \mathbb{E}\widehat{\gamma}_{k\cdot}\right)^\top Z_X^\top Z_X\left(\widehat{\gamma}_{k\cdot} - \mathbb{E}\widehat{\gamma}_{k\cdot}\right) \\ =\ &\left(\widehat{\gamma}_{k\cdot} - \mathbb{E}\widehat{\gamma}_{k\cdot}\right)^\top D_k^{-1/2}D_k^{1/2}Z_X^\top Z_X D_k^{1/2}D_k^{-1/2}\left(\widehat{\gamma}_{k\cdot} - \mathbb{E}\widehat{\gamma}_{k\cdot}\right) \\ =\ &\left(D_k^{-1/2}(\widehat{\gamma}_{k\cdot} - \mathbb{E}\widehat{\gamma}_{k\cdot})\right)^\top\left(\sum_{s=1}^{p}\eta_{ks}\mathbf{v}_{ks}\mathbf{v}_{ks}^\top\right)\left(D_k^{-1/2}(\widehat{\gamma}_{k\cdot} - \mathbb{E}\widehat{\gamma}_{k\cdot})\right) \\ =\ &\sum_{s=1}^{p}\eta_{ks}\left(v_{ks}^\top D_k^{-1/2}(\widehat{\gamma}_{k\cdot} - \mathbb{E}\widehat{\gamma}_{k\cdot})\right)^2.\end{aligned}$$

Here we use the eigendecomposition

$$D_k^{1/2}Z_X^\top Z_X D_k^{1/2} = \sum_{s=1}^{p}\eta_{ks}v_{ks}v_{ks}^\top \in \mathbb{R}^{n\times n}$$

where for $s = 1, \ldots, p$; η_{ks} are the eigenvalues and $v_{ks} \in \mathbb{R}^n$ are mutually orthonormal eigenvectors.

Finally, under H_0, we have from (35) that

$$\left(v_{ks}^\top D_k^{-1/2}(\widehat{\gamma}_{k\cdot} - \mathbb{E}\widehat{\gamma}_{k\cdot})\right)^2 \overset{iid}{\sim} \chi_1^2, \text{ for } k = 1, \ldots, d,$$

conditional on $Y_{ij}^{(1)}$. We note here that the same conclusion is asymptotically correct if the eigenvectors v_{ks} are incoherent.

Moreover, if $\sum_{i=1}^{n} m_j \gg d$, it follows by consistency of sample eigenvectors that there exists an eigendecomposition $\Sigma_0 = \sum_k \lambda_k u_k u_k^\top$ such that the following holds. Let us write $\widehat{\Sigma}_0 := \frac{1}{\sum m_i}\sum_{ij} Y_{ij}^{(1)}(Y_{ij}^{(1)})^\top$. Then by covariance matrix concentration results it follows

that

$$\left\|\widehat{\Sigma}_0 - \Sigma_0\right\| = O_p\left(\sqrt{\frac{d}{\sum_i m_i}}\right).$$

By Weyl's theorem, we then have $\widehat{\lambda}_k = \widehat{u}_k^\top \widehat{\Sigma}_0 \widehat{u}_k$ satisfies

$$|\widehat{\lambda}_k - \lambda_k| = O_p\left(\sqrt{\frac{d}{\sum_i m_i}}\right).$$

and hence

$$\left|\widehat{u}_k^\top \Sigma_0 \widehat{u}_k - \lambda_k\right| \leq \left|\widehat{u}_k^\top (\Sigma_0 - \widehat{\Sigma}_0)\widehat{u}_k\right| + \left|\widehat{\lambda}_k - \lambda_k\right| = O_p\left(\sqrt{\frac{d}{\sum_i m_i}}\right). \tag{36}$$

Since two square matrices A and B have the same eigenvalues for AB and BA, note that η_{ks} are the eigenvalues of $D_k Z_X^\top Z_X$. Now similarly let η_{ks}^* be the eigenvalues of $D_k^* Z_X^\top Z_X$. Here

$$D_k^* = \operatorname{diag}\left(\left\{\frac{\lambda_k}{2m_1} \cdots \frac{\lambda_k}{2m_n}\right\}\right).$$

It follows from (36) that

$$\eta_{ks} = \eta_{ks}^* \left\{1 + O_p\left(\sqrt{\frac{d}{\sum_i m_i}}\right)\right\}.$$

Let us define a random variable $\tilde{F}_* = \sum_{k=1}^d \sum_{s=1}^p \eta_{ks}^* (\chi_1^2)_{ks}$ where $(\chi_1^2)_{ks}$ for $k = 1, \ldots, d$ and $s = 1, \ldots, p$ denote dp i.i.d. χ_1^2 random variables. By the previous equation it follows that there exists a constant $C > 0$ such that

$$\mathbb{P}\left(\widehat{F}_* \leq x\right) \in \left[\mathbb{P}\left\{\tilde{F}_* \leq x\left(1 - C\sqrt{\frac{d}{\sum_i m_i}}\right)\right\}, \mathbb{P}\left\{\tilde{F}_* \leq x\left(1 + C\sqrt{\frac{d}{\sum_i m_i}}\right)\right\}\right] \text{ for all } x \in \mathbb{R}. \tag{37}$$

Proof of Proposition 2. Note that for $1 \leq s, t \leq n$, we have

$$\left(\mathbf{Z}_X^\top \mathbf{Z}_X - \mathbf{Z}_{X^{(1)}}^\top \mathbf{Z}_{X^{(1)}}\right)_{st} = (X_s - \bar{X})^\top \Sigma_X^{-1} (X_t - \bar{X}) - (X_s^{(1)} - \bar{X}^{(1)})^\top \Sigma_{X^{(1)}}^{-1} (X_t^{(1)} - \bar{X}^{(1)})$$

By block matrix inversion lemma, we have

$$\Sigma_X^{-1} = \begin{pmatrix} \Sigma_{X^{(1)}}^{-1} + \Sigma_{X^{(1)}}^{-1}\Sigma_{12}\Sigma_{22|1}^{-1}\Sigma_{12}^\top\Sigma_{X^{(1)}}^{-1} & -\Sigma_{X^{(1)}}^{-1}\Sigma_{12}\Sigma_{22|1}^{-1} \\ -\Sigma_{22|1}^{-1}\Sigma_{12}^\top\Sigma_{X^{(1)}}^{-1} & \Sigma_{22|1}^{-1} \end{pmatrix}.$$

Thus, for $1 \leq s, t \leq n$, we obtain

$$\begin{aligned}
&\left(Z_X^\top Z_X - Z_{X^{(1)}}^\top Z_{X^{(1)}}\right)_{st} \\
&= (X_s^{(1)} - \bar{X}^{(1)})^\top \Sigma_{X^{(1)}}^{-1}\Sigma_{12}\Sigma_{22|1}^{-1}\Sigma_{12}^\top\Sigma_{X^{(1)}}^{-1}(X_t^{(1)} - \bar{X}^{(1)}) \\
&\quad - (X_s^{(1)} - \bar{X}^{(1)})^\top \Sigma_{X^{(1)}}^{-1}\Sigma_{12}\Sigma_{22|1}^{-1}(X_t^{(2)} - \bar{X}^{(2)}) \\
&\quad - (X_s^{(2)} - \bar{X}^{(2)})^\top \Sigma_{22|1}^{-1}\Sigma_{12}^\top\Sigma_{X^{(1)}}^{-1}(X_t^{(1)} - \bar{X}^{(1)}) + (X_s^{(2)} - \bar{X}^{(2)})\Sigma_{22|1}^{-1}(X_t^{(2)} - \bar{X}^{(2)}) \\
&= \left(\Sigma_{12}^\top\Sigma_{X^{(1)}}^{-1}(X_s^{(1)} - \bar{X}^{(1)}) - (X_s^{(2)} - \bar{X}^{(2)})\right)^\top \Sigma_{22|1}^{-1}\left(\Sigma_{12}^\top\Sigma_{X^{(1)}}^{-1}(X_t^{(1)} - \bar{X}^{(1)}) - (X_t^{(2)} - \bar{X}^{(2)})\right).
\end{aligned}$$

In particular, we can rewrite

$$\left(Z_X^\top Z_X - Z_{X^{(1)}}^\top Z_{X^{(1)}}\right) = Z_{X^{(2|1)}}^\top Z_{X^{(2|1)}}$$

where $Z_{X^{(2|1)}}$ is as defined in the statement of the proposition. The conclusion now follows by the definition of $\widehat{F}_P^*$ from equation (25). □

References

Álvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J. & Matrán, C. (2016). A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications* **441**, 744–762.

Chewi, S., Maunu, T., Rigollet, P. & Stromme, A. J. (2020). Gradient descent algorithms for bures-wasserstein barycenters. In *Conference on Learning Theory*. PMLR.

Davinelli, S., Willcox, D. & Scapagnini, G. (2012). Extending healthy ageing: nutrient sensitive pathway and centenarian population. *Immunity & Ageing* **9**, 9.

de Lucia, C., Murphy, T., Steves, C. & et al. (2020). Lifestyle mediates the role of nutrient-sensing pathways in cognitive aging: cellular and epidemiological evidence. *Communications Biology* **3**, 157.

Elyahu, Y., I, H., I, E.-M., O, B., I, S., M, S., K, M., A, N., E, E., A, V., E, S., V, C.-C., N, F., E, Y.-L. & A, M. (2019). Aging promotes reorganization of the cd4 t cell landscape toward extreme regulatory and effector phenotypes. *Science Advances* **5**, eaaw8330.

Flury, B. N. (1984). Common principal components in k groups. *Journal of the American Statistical Association* **79**, 892–898.

Flury, B. N. (1986). Asymptotic theory for common principal component analysis. *The annals of Statistics* , 418–430.

Givens, C. & Shortt, R. (1984). A class of wasserstein metrics for probability distributions. *Michigan Math. J.* **31**, 231–240.

Harris, B., Crow, M., Fischer, S. & Gillis, J. (2021). Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain. *Cell Systems* **12**, 748–756.e3.

Hoff, P. D. & Niu, X. (2012). A covariance regression model. *Statistica Sinica* , 729–753.

Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F. & Luo, Y. (2022). Single-cell rna sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine* **12**.

Knott, M. & Smith, C. S. (1984). On the optimal mapping of distributions. *J. Optim. Theory Appl.* **43**, 39–49.

Petersen, A., Liu, X. & Divani, A. A. (2021). Wasserstein F-tests and confidence bands for the Fréchet regression of density response curves. *The Annals of Statistics* **49**, 590 – 611.

Petersen, A. & Müller, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics* **47**, 691 – 719.

Ribeiro, D., Ziyani, D. & Delaneau, O. (2022). Shared regulation and functional relevance of local gene co-expression revealed by single cell analysis. *Communications Biololgy* **5**.

Tomokazu, S. & Hafler, D. (2022). Population genetics meets single-cell sequencing. *Science* **376**, 134–135.

Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M. G., Andersen, S., Lu, Q., Rowson, A., Taylor, T. R., Clarke, L. et al. (2022). Single-cell eqtl mapping identifies cell type–specific genetic control of autoimmune disease. *Science* **376**, eabf3041.

Zou, T., Lan, W., Wang, H. & Tsai, C.-L. (2017). Covariance regression analysis. *Journal of the American Statistical Association* **112**, 266–281.