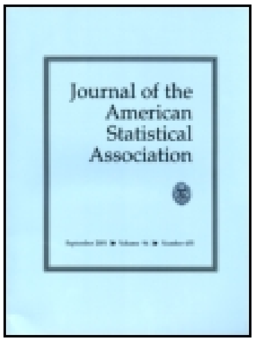

Journal of the American Statistical Association

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

High-Dimensional MANOVA Via Bootstrapping and Its Application to Functional and Sparse Count Data

Zhenhua Lin, Miles E. Lopes & Hans-Georg Müller

To cite this article: Zhenhua Lin, Miles E. Lopes & Hans-Georg Müller (2021): High-Dimensional MANOVA Via Bootstrapping and Its Application to Functional and Sparse Count Data, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.1920959

To link to this article: https://doi.org/10.1080/01621459.2021.1920959

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–15: Theory and Methods
https://doi.org/10.1080/01621459.2021.1920959

High-Dimensional MANOVA Via Bootstrapping and Its Application to Functional and Sparse Count Data

Zhenhua Lin[a], Miles E. Lopes[b], and Hans-Georg Müller[c]

[a]Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore; [b]Department of Statistics, University of California, Davis, CA; [c]Department of Statistics, University of California, Davis, CA

ABSTRACT
We propose a new approach to the problem of high-dimensional multivariate ANOVA via bootstrapping max statistics that involve the differences of sample mean vectors. The proposed method proceeds via the construction of simultaneous confidence regions for the differences of population mean vectors. It is suited to simultaneously test the equality of several pairs of mean vectors of potentially more than two populations. By exploiting the variance decay property that is a natural feature in relevant applications, we are able to provide dimension-free and nearly parametric convergence rates for Gaussian approximation, bootstrap approximation, and the size of the test. We demonstrate the proposed approach with ANOVA problems for functional data and sparse count data. The proposed methodology is shown to work well in simulations and several real data applications.

ARTICLE HISTORY
Received Aug 2020
Accepted April 2021

KEYWORDS
Bootstrap methods; Functional data analysis; Gaussian approximation; Hypothesis testing; Mean function; Physical activity; Poisson data; Simultaneous confidence intervals

1. Introduction

The MANOVA problem of detecting significant differences among the means of multivariate populations is of central importance in a myriad of statistical applications. However, the classical MANOVA approaches are only intended to handle low-dimensional settings where the number of covariates is much smaller than the sample size, which is a crucial limitation for modern high-dimensional data analysis. Due to the demand for methodology that provides valid inference for high-dimensional data, the challenge of finding suitable new MANOVA methods has developed into a major line of research. For example, the special case of high-dimensional two-sample testing has been investigated by Bai and Saranadasa (1996), Lopes, Jacob, and Wainwright (2011), Cai, Liu, and Xia (2014), Thulin (2014), Xu et al. (2016), Zhang and Pan (2016), and Zhang et al. (2019b) under the condition that populations share a common covariance matrix, while procedures designed by Chen and Qin (2010), Feng and Sun (2015), Feng et al. (2015), Gregory et al. (2015), Städler and Mukherjee (2016), Chang et al. (2017), and Xue and Yao (2020) do not require such a common covariance assumption. For the more general multiple-sample problem, methods and theory were studied by Fujikoshi, Himeno, and Wakaki (2004), Srivastava and Fujikoshi (2006), Schott (2007), Yamada and Srivastava (2012), Srivastava and Kubokawa (2013), Cai and Xia (2014), Zhang, Guo, and Zhou (2017), Bai, Choi, and Fujikoshi (2018), and Li, Aue, and Paul (2020) when the populations share common covariance structure, while Zhang and Xu (2009), Yamada and Himeno (2015), Li et al. (2017), Hu et al. (2017), Zhou, Guo, and Zhang (2017), and Zhang et al. (2018) eliminated the requirement of common covariance. Among these, Chang et al. (2017), Zhang et al. (2018), and Xue and Yao (2020) adopted a bootstrap approach following Chernozhukov, Chetverikov, and Kato (2013) and Chernozhukov, Chetverikov, and Kato (2017).

An important observation in this context is that the variances of variables often exhibit a certain decay pattern. As an example, consider a multinomial model of p categories. Without loss of generality, assume that the probabilities of the p categories are ordered as $\pi_1 \geq \cdots \geq \pi_p$. Since the probabilities sum to one, it follows that the variance $\sigma_j^2 = \pi_j(1-\pi_j)$ of the jth category must decay at least as fast as j^{-1}. Additional examples that arise in connection with principal component analysis and the Fourier coefficients of functional data may be found in Lopes, Lin, and Müller (2020).

When the structure of variance decay is available, Lopes, Lin, and Müller (2020) showed that near-parametric and dimension-free rates of Gaussian and bootstrap approximation can be established for max statistics of the form $\max_{1\leq j\leq p}\sqrt{n}\{\bar{X}-\mu\}(j)/\sigma_j^\tau$. In this expression, $\bar{X} = (\bar{X}(1),\ldots,\bar{X}(p))$ is the sample mean of n independent and identically distributed random vectors with mean vector $\mu = (\mu(1),\ldots,\mu(p))$ and coordinate-wise variances $\sigma_1^2,\ldots,\sigma_p^2$, while the symbol τ denotes a tuning parameter in the interval $[0,1)$. Remarkably, the near-parametric rates of approximation remain valid even when the decay is very weak, that is, $\sigma_j \asymp j^{-\alpha}$ for an arbitrarily small $\alpha > 0$. In this paper, we harness such decay patterns to develop promising bootstrap-based inference for the high-dimensional MANOVA problem.

We consider a general setting with $K \geq 2$ populations having mean vectors $\mu_1,\ldots,\mu_K \in \mathbb{R}^p$. For any collection of ordered pairs $\mathcal{P}$ taken from the set $\{(k,l) : 1 \leq k < l \leq K\}$, the

CONTACT Zhenhua Lin linz@nus.edu.sg Department of Statistics and Applied Probability, National University of Singapore, Singapore 119260, Singapore.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

hypothesis testing problem of interest is

$$\begin{aligned} &\mathrm{H}_0: \mu_k = \mu_l \quad \text{for all } \ (k,l) \in \mathcal{P} \qquad \text{versus} \\ &\mathrm{H}_\mathrm{a}: \mu_k \neq \mu_l \quad \text{for some} \quad (k,l) \in \mathcal{P}. \end{aligned} \tag{1}$$

Note that this includes a very general class of null hypotheses of possible interest. The proposed strategy is to construct simultaneous confidence region (SCR) for the differences $\mu_k - \mu_l$ for all pairs in $\mathcal{P}$ via bootstrapping a maximum-type statistic related to $\mu_k - \mu_l$ across all coordinates and all pairs. In addition, we adopt the idea of partial standardization developed in Lopes, Lin, and Müller (2020) to take advantage of the variance decay. This differs from the existing bootstrap-based methods proposed in Chang et al. (2017), Xue and Yao (2020), and Zhang et al. (2018) that do not exploit the decay. Furthermore, in the first two papers the authors consider only one- or two-sample problems, and in the last article only the standard global null hypothesis $\mu_1 = \cdots = \mu_K$.

The proposed method has several favorable properties:

- There is flexibility in the choice of null hypothesis. In addition to the basic global null hypothesis $\mu_1 = \cdots = \mu_K$, which corresponds to choosing $\mathcal{P} = \{(k,l) : 1 \le k < l \le K\}$, we can also test more specific hypotheses. For instance, the null hypothesis $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$ corresponds to $\mathcal{P} = \{(1,2),(3,4)\}$. In general, whenever $\mathcal{P}$ contains more than one pair, traditional methods often require that two or more separate tests are performed. This requires extra adjustments for multiple comparisons, which often have a negative impact on power. Indeed, the effect of multiplicity can be severe, because the number of pairs $|\mathcal{P}|$ may grow quadratically as a function of K, as in the case of the global null hypothesis with $|\mathcal{P}| = K(K-1)/2$.
- The proposed method performs the test via constructing SCR for the differences $\mu_k - \mu_l$ indexed by $(k,l) \in \mathcal{P}$. Such SCR are also valuable in their own right (in addition to their utility for hypothesis testing), as they provide quantitative information about the separation of the mean vectors $\mu_1, \dots, \mu_K$ that is often of interest in applications.
- When the null hypothesis is rejected, the proposed approach makes it possible to immediately identify pairs of populations that have significantly different means without performing additional tests. By contrast, additional testing is often necessary when one adopts and extends traditional MANOVA approaches.
- Like Chang et al. (2017), Zhang et al. (2018), and Xue and Yao (2020), who essentially proposed two-sample or multiple-sample comparisons based on bootstrapping, we do not require that the ratio of the sample sizes of any pair of populations converges to a specific limit.
- In contrast to the testing procedures of Chang et al. (2017), Zhang et al. (2018) (where the convergence rates for the size of the test are not established), and the method of Xue and Yao (2020) (for which the convergence rate is at most $\sqrt{\log p}/n^{1/6}$), the proposed approach is shown to enjoy a near-parametric rate of convergence. Furthermore, this near-parametric rate is free of the dimension p and holds under mild assumptions. These improvements are achieved by exploiting variance decay.

To demonstrate the usefulness of the proposed approach, we apply our procedure to perform ANOVA for functional data and sparse count data. Functional data are commonly encountered in many types of statistical analysis, as surveyed in the monographs (Ramsay and Silverman 2005; Ferraty and Vieu 2006; Horváth and Kokoszka 2012; Zhang 2013; Hsing and Eubank 2015; Kokoszka and Reimherr 2017) and review papers (Wang, Chiou, and Müller 2016; Aneiros et al. 2019). Previous examples of methods for functional ANOVA are pointwise F-tests (Ramsay and Silverman 2005, p. 227), an integrated F-test and its variants (Shen and Faraway 2004; Zhang 2011, 2013), globalization of pointwise F-tests (Zhang and Liang 2014), a test based on the maximum of pointwise F-statistics (Zhang et al. 2019a), the HANOVA method (Fan and Lin 1998), L^2 norm-based methods (Faraway 1997; Zhang and Chen 2007), random projection-based test (Cuesta-Albertos and Febrero-Bande 2010), a GET with graphical interpretation (Mrkvička et al. 2020), and an empirical likelihood ratio approach (Chang and McKeague 2020), in addition to resampling methods (Zhang 2013; Paparoditis and Sapatinas 2016).

While the proposed approach makes use of the techniques and some results developed in Lopes, Lin, and Müller (2020), adapting these results to the multiple-sample setting is a major challenge. The key obstacle is that, in contrast to the situation studied in Lopes, Lin, and Müller (2020), the max statistic (2) in the MANOVA setting is not the maximum of an average of independent vectors. Overcoming this difficulty requires a delicate transformation of the statistic to represent it as the maximum of the average of independent random vectors that are further transformations of the data; see Proposition A.1 in the supplementary material. In addition, the theory here is more comprehensive in the way that it accounts for the effect using estimated standard deviations $\hat{\sigma}_j$ in the SCR. This is done by establishing a uniform bound on the estimation error of $\hat{\sigma}_j$ over all coordinates and groups, which holds when the data satisfy a basic continuity assumption; see Lemma E.10 in the supplementary material.

The rest of the article is structured as follows. In Section 2, we present the details of the proposed method. In Section 3, we establish theoretical guarantees for bootstrapping max statistics under a multiple-sample setting, including a result on the convergence rate of the empirical size of the proposed test. Our signature application to functional ANOVA is given in Section 4 and the second application to sparse count data is given in Section 5. We conclude the article in Section 6.

2. High-Dimensional Multiple-Sample Test

Consider K independent groups of observations, where we assume that for the kth group one has n_k iid (independently and identically distributed) p-dimensional observations $X_{k,1}, \dots, X_{k,n_k}$ with mean $\mu_k \in \mathbb{R}^p$. Our goal is to test any of the null hypotheses in Equation (1) based on these data.

To motivate our approach, consider a two-sample test in the classical setting that corresponds to the special case $p = 1$ and $K = 2$ with $(k,l) = (1,2)$. The common statistic $T =$

$\{(\bar{X}_k - \mu_k) - (\bar{X}_l - \mu_l)\}/\sqrt{\mathrm{var}(\bar{X}_k - \bar{X}_l)}$ asymptotically follows a standard Gaussian distribution, where $\bar{X}_k = n_k^{-1}\sum_{i=1}^{n_k} X_{k,i}$ denotes the sample mean of the kth group for $k = 1, 2$. This statistic can be used to construct a confidence interval of level $1 - \varrho$ for the difference $\mu_k - \mu_l$, which can then be used to implement the standard two-sample test at the level ϱ. When $p > 1$, one can construct a SCR for $\mu_k - \mu_l \in \mathbb{R}^p$ in terms of the distribution of the max statistic

$$M'(k,l) = \max_{1\le j\le p} \frac{\{\bar{X}_k(j) - \mu_k(j)\} - \{\bar{X}_l(j) - \mu_l(j)\}}{\sqrt{\mathrm{var}(\bar{X}_k(j) - \bar{X}_l(j))}}.$$

For the general case when $K \ge 2$, it is natural to consider the max statistic $M' = \max_{(k,l)\in\mathcal{P}} M'(k,l)$. One may equivalently rewrite the statistic $M'(k,l)$ as

$$M'(k,l) = \max_{1\le j\le p}\left(\sqrt{\tfrac{n_l}{n_k+n_l}}\,\frac{S_{k,j}}{\sigma_{k,l,j}} - \sqrt{\tfrac{n_k}{n_k+n_l}}\,\frac{S_{l,j}}{\sigma_{k,l,j}}\right),$$

where $S_k = n_k^{-1/2}\sum_{i=1}^{n_k}(X_{k,i} - \mu_k)$, $S_{k,j} = S_k(j)$ denotes the jth coordinate, and $\sigma^2_{k,l,j} = \{n_l \mathrm{var}(X_k(j)) + n_k \mathrm{var}(X_l(j))\}/(n_k + n_l)$. As shown in Lopes, Lin, and Müller (2020), when the variances $\sigma^2_{k,l,j}$ exhibit a decay pattern, it is beneficial to use *partial standardization*,

$$M(k,l) = \max_{1\le j\le p}\left(\sqrt{\tfrac{n_l}{n_k+n_l}}\,\frac{S_{k,j}}{\sigma^\tau_{k,l,j}} - \sqrt{\tfrac{n_k}{n_k+n_l}}\,\frac{S_{l,j}}{\sigma^\tau_{k,l,j}}\right) \quad \text{and} \qquad (2)$$

$$M = \max_{(k,l)\in\mathcal{P}} M(k,l),$$

where $\tau \in [0, 1)$ is a parameter that may be tuned to maximize power.

Remark. To intuitively understand the role of τ, it is helpful to consider the extreme cases of $\tau = 1$ (ordinary standardization) and $\tau = 0$ (no standardization). In the case of $\tau = 1$, the jth difference in Equation (2) has variance equal to 1 for every $j = 1, \dots, p$, and hence, the "low-dimensional structure" of variance decay is eliminated. Likewise, in this situation, all of the p coordinates are "equally important," which makes the problem genuinely high-dimensional—and hence, makes bootstrap approximation more difficult. In the opposite case when $\tau = 0$, a different issue arises. It can be seen from Equation (3) below that all of the p SCRs will have the same width. This is undesirable, as the widths of the intervals should be adapted to the variance of each coordinate. In view of these undesirable effects when choosing the endpoints $\tau = 1$ or $\tau = 0$, the proposed partial standardization seeks a tradeoff by allowing for intermediate values of τ between 0 and 1.

As M is the maximum of random variables that are in turn coordinate-wise maxima of a random vector, it is difficult to derive its distribution.[1] This difficulty, fortunately, can be circumvented efficiently by bootstrapping, as follows. Let $\hat{\Sigma}_k = n_k^{-1}\sum_{i=1}^{n_k}(X_{k,i} - \bar{X}_k)(X_{k,i} - \bar{X}_k)^\top$ be the sample covariance of the kth group. Define the bootstrap version of S_k by $S^\star_k \sim N(0, \hat{\Sigma}_k)$. (An equivalent definition is $S^\star_k = n_k^{-1/2}\sum_{i=1}^{n_k} X^\star_{k,i}$ with $X^\star_{k,i}$ iid sampled from $N(0, \hat{\Sigma}_k)$.) Likewise, the bootstrap version of $M(k,l)$ is defined by

$$M^\star(k,l) = \max_{1\le j\le p}\left(\sqrt{\tfrac{n_l}{n_k+n_l}}\,\frac{S^\star_{k,j}}{\hat{\sigma}^\tau_{k,l,j}} - \sqrt{\tfrac{n_k}{n_k+n_l}}\,\frac{S^\star_{l,j}}{\hat{\sigma}^\tau_{k,l,j}}\right),$$

where $\hat{\sigma}^2_{k,l,j}$ are diagonal elements of $\hat{\Sigma}_{k,l} = \frac{n_l}{n_k+n_l}\hat{\Sigma}_k + \frac{n_k}{n_k+n_l}\hat{\Sigma}_l$, and altogether, the bootstrap version of M is defined by

$$M^\star = \max_{(k,l)\in\mathcal{P}} M^\star(k,l).$$

For a given dataset $X = \{X_{k,i} : 1 \le k \le K, 1 \le i \le n_k\}$, we generate $B \ge 1$ independent samples of $(S^\star_1, \dots, S^\star_K)$, which yield B independent samples of $M^\star$. Then, the empirical quantile function of these samples of $M^\star$, denoted by $\hat{q}_M(\cdot)$, serves as an estimate of the quantile function $q_M(\cdot)$ of M.

Analogously, we define the min statistic

$$L(k,l) = \min_{1\le j\le p}\left(\sqrt{\tfrac{n_l}{n_k+n_l}}\,\frac{S_{k,j}}{\sigma^\tau_{k,l,j}} - \sqrt{\tfrac{n_k}{n_k+n_l}}\,\frac{S_{l,j}}{\sigma^\tau_{k,l,j}}\right) \quad \text{and}$$

$$L = \min_{(k,l)\in\mathcal{P}} L(k,l),$$

as well as their bootstrap counterparts,

$$L^\star(k,l) = \min_{1\le j\le p}\left(\sqrt{\tfrac{n_l}{n_k+n_l}}\,\frac{S^\star_{k,j}}{\hat{\sigma}^\tau_{k,l,j}} - \sqrt{\tfrac{n_k}{n_k+n_l}}\,\frac{S^\star_{l,j}}{\hat{\sigma}^\tau_{k,l,j}}\right) \quad \text{and}$$

$$L^\star = \min_{(k,l)\in\mathcal{P}} L^\star(k,l).$$

Similarly, the quantile function of $L^\star$ can be obtained by drawing samples from the distributions $N(0, \hat{\Sigma}_k)$.

Finally, the $1 - \varrho$ two-sided SCRs for the jth coordinates of $\mu_k - \mu_l$ for $j = 1, \dots, p$, $(k,l) \in \mathcal{P}$, are given by

$$\mathrm{SCR}(k,l,j) = \left[\bar{X}_k(j) - \bar{X}_l(j) - \frac{\hat{q}_M(1-\varrho/2)\hat{\sigma}^\tau_{k,l,j}}{\sqrt{n_{k,l}}},\right.$$

$$\left.\bar{X}_k(j) - \bar{X}_l(j) - \frac{\hat{q}_L(\varrho/2)\hat{\sigma}^\tau_{k,l,j}}{\sqrt{n_{k,l}}}\right], \qquad (3)$$

where $n_{k,l} := n_k n_l/(n_k + n_l)$ denotes the harmonic sample size of the kth and lth groups. With these SCRs in hand, we perform the test in (1) by rejecting the null hypothesis at the significance level ϱ if $0 \notin \mathrm{SCR}(k,l,j)$ for some $(k,l) \in \mathcal{P}$ and $j = 1, \dots, p$. One-sided SCRs can be constructed and one-sided hypothesis tests can be conducted in a similar fashion. For the testing problem (1), it is often desirable to obtain the p-value, which corresponds to the largest value of ϱ such that all SCRs in Equation (3) contain zero and can easily be found numerically.

In practical applications, one needs to determine a value for the parameter τ. Although in the next section, it is shown that any fixed value in $[0, 1)$ gives rise to the same asymptotic behavior of the proposed test, a data-driven method to optimize the empirical power is desirable. We propose to select the value of τ that yields the smallest p-value while keeping the size at the nominal level ϱ. We first observe that for a given value of τ, the above bootstrap test provides a corresponding p-value. It remains to estimate the empirical size for a given value of τ. To this end, we propose the following resampling approach.

[1] Note that M itself is not a test statistic since it involves unknown parameters, but being able to estimate the quantiles of M will enable our testing procedure based on SCRs.

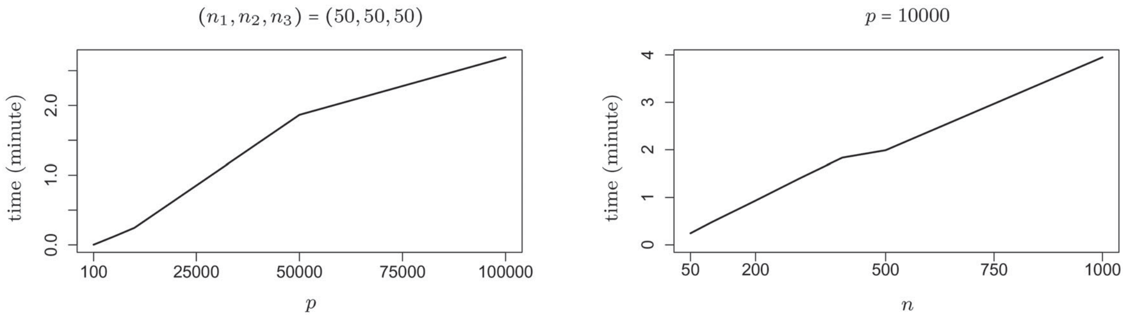

Figure 1. Computation time in a server with an NVIDIA Quadro P400 graphics card.

First, the data are centered within each group, so that the null hypothesis holds for the centered data. For each group, a new sample of the same size is generated by resampling the original dataset with replacement. Then, the proposed test is applied on the new samples with the nominal significance level ϱ. This process is repeated several times, for example, 100 times, and the empirical size is estimated by the proportion of the resampled datasets that lead to rejecting the null hypothesis. If a value of τ yields an empirical size that is bounded by the nominal level ϱ, then it is retained, and from these retained values of τ, the one corresponding to the smallest p-value is selected.

To tackle the additional computational burden that this incurs, one can leverage the two levels of parallelism of the proposed algorithm: Each candidate value of τ in a grid can be examined in parallel, and for a given τ, all the subsequent computations are parallel. Therefore, the proposed method is scalable with modern cloud, cluster or GPU (graphics processing unit) based computing. For illustration, we created an R software to implement the above parallel algorithm for a GPU based platform. Figure 1 shows the computation time that includes selecting a value for τ from 11 candidate values, constructing the SCRs and performing the test, for datasets of $K = 3$ groups, $(n_1, n_2, n_3) = (n, n, n)$ samples and p dimensions. It is observed that the computation time scales efficiently in both n and p.

3. Theory

3.1. Bootstrapping Max Statistics for Multiple Samples

Notation. The identity matrix of size $p \times p$ is denoted by I_p. For a deterministic vector $v \in \mathbb{R}^p$ and $r > 0$, let $\|v\|_r = (\sum_{j=1}^{p} |v_j|^r)^{1/r}$, and for a scalar random variable ξ, let $\|\xi\|_r = \mathbb{E}(|\xi|^r)^{1/r}$. The ψ_1-Orlicz norm of a random variable ξ is denoted and defined by $\|\xi\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}[\exp(|\xi|/t)] \leq 2\}$. If a and b are real numbers, then we write $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

Conventions. The main results are formulated in terms of a sequence of models indexed by the integer $n = \min\{n_1, \ldots, n_K\}$. All aspects of these models may depend on n, except where stated otherwise. Likewise, the following numbers may depend on n: the dimension p, the number of groups K, the group sample sizes $n_1, \ldots, n_K$,[2] and the tuning parameter τ. The set of pairs $\mathcal{P}$, as well as the population distributions of the groups may also depend on n. Accordingly, if it is stated that a constant c does not depend on n, then it is understood that c does not depend on any of these other numbers or objects. For constants of this type, the symbol c will often be reused with different values at each occurrence. If a_n and b_n are two sequences of nonnegative real numbers, then $a_n \lesssim b_n$ means that there is a constant $c > 0$ not depending on n, such that $a_n \leq c b_n$ holds for all large n. If both of the conditions $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold, then we write $a_n \asymp b_n$.

Assumption 1 (Data-generating model).

(i) For each $k \in \{1, \ldots, K\}$, there exists a vector $\mu_k \in \mathbb{R}^p$ and a positive semidefinite matrix $\Sigma_k \in \mathbb{R}^{p \times p}$, such that the observations $X_{k,1}, \ldots X_{k,n_k} \in \mathbb{R}^p$ are generated as $X_{k,i} = \mu_k + \Sigma_k^{1/2} Z_{k,i}$ for each $1 \leq i \leq n_k$, where $Z_{k,1}, \ldots, Z_{k,n_k} \in \mathbb{R}^p$ are iid random vectors.

(ii) There is a constant $c_0 > 0$ not depending on n, such that for each $k \in \{1, \ldots, K\}$, the random vector $Z_{k,1}$ satisfies $\sup_{\|u\|_2=1} \|Z_{k,1}^\top u\|_{\psi_1} \leq c_0$, as well as $\mathbb{E}Z_{k,1} = 0$ and $\mathbb{E}(Z_{k,1}Z_{k,1}^\top) = I_p$.

In the above assumption, the mean vectors μ_k and covariance matrices Σ_k are allowed to vary with the sample size n_k. Also, the random vectors $Z_{1,1}, \ldots, Z_{1,n_1}, \ldots, Z_{K,1}, \ldots, Z_{K,n_K}$ across different populations are independent, and $Z_{1,1}, \ldots, Z_{K,1}$ may have different distributions.

To state the next assumption, for $d \in \{1, \ldots, p\}$, we use $\mathcal{J}_k(d)$ to denote a set of indices corresponding to the d largest values among $\sigma_{k,1}, \ldots, \sigma_{k,p}$. In addition, let $R_k(d) \in \mathbb{R}^{d \times d}$ denote the correlation matrix of the random variables $\{X_{k,1}(j) : j \in \mathcal{J}_k(d)\}$. Lastly, let $a \in (0, 1/2)$ be a fixed constant, and define the integers ℓ_k and m_k according to

$$\ell_k = \lceil (1 \vee \log^3 n_k) \wedge p \rceil,$$

$$m_k = \lceil (\ell_k \vee n_k^{\frac{1}{\log(n_k)^a}}) \wedge p \rceil.$$

Assumption 2 (Structural assumptions).

(i) The parameters $\sigma_{k,1}, \ldots, \sigma_{k,p}$ are positive, and there are positive constants α, c_1, and $c_\circ \in (0, 1)$, not depending on

[2] that is, for each $n = 1, 2, \ldots$, the equation $n = \min\{n_1(n), \ldots, n_K(n)\}$ is satisfied.

n, such that for each $k \in \{1, \dots, K\}$,

$$\sigma_{k,(j)} \le c_1 j^{-\alpha} \quad \text{for all } j \in \{m_k, \dots, p\},$$

$$\sigma_{k,(j)} \ge c_\circ j^{-\alpha} \quad \text{for all } j \in \{1, \dots, m_k\},$$

where $\sigma_{k,(j)}$ denotes the jth largest value of $\sigma_{k,1}, \dots, \sigma_{k,p}$.

(ii) There exists a constant $\epsilon_0 \in (0, 1)$, not depending on n, such that for $k = 1, \dots, K$,

$$\max_{i \ne j} R_{k,i,j}(\ell_k) \le 1 - \epsilon_0,$$

where $R_{k,i,j}(\ell_k)$ denotes the (i,j) entry of the matrix $R_k(\ell_k)$. Also, for $k = 1, \dots, K$, the matrix $R_k^+(\ell_k)$ with (i,j) entry given by $\max\{R_{k,i,j}(\ell_k), 0\}$ is positive semidefinite. Moreover, there is a constant $C_0 > 0$, not depending on n, such that for each $k = 1, \dots, K$, we have

$$\sum_{1 \le i < j \le \ell_k} R^+_{k,i,j}(\ell_k) \le C_0 \ell_k.$$

The above two assumptions are multiple-sample analogs of assumptions in Lopes, Lin, and Müller (2020), where examples of correlation matrices satisfying the above conditions are given. The following assumption imposes constraints on τ in conjunction with n and on the sample sizes $n_1, \dots, n_K$.

Assumption 3. There exist positive constants c_2 and c_3 not depending on n such that the bounds $c_2 \le \frac{n_k}{n_k + n_l} \le c_3$ hold for all $k, l \in \{1, \dots, K\}$. Also, the conditions $(1 - \tau)\sqrt{\log n} \gtrsim 1$ and $\max\{K, |\mathcal{P}|\} \lesssim e^{\sqrt{\log n}}$ hold.

In the last assumption, note that τ is allowed to approach to 1 at a slow rate. Although $n_1, \dots, n_K$ are required to be of the same order, their ratios do not have to converge to certain limits. Such convergence conditions are required by some of the test procedures surveyed in Section 1 based on asymptotic limit distributions of test statistics rather than bootstrap. Also, it is notable that the current setting allows $K \to \infty$ and $|\mathcal{P}| \to \infty$ as $n \to \infty$. Overall, Assumptions 1–3 are quite mild and are satisfied for many relevant applications, with examples in Sections 4 and 5.

Let $\tilde{S}_k \sim N(0, \Sigma_k)$ for each $k = 1, \dots, K$, and define the Gaussian counterparts of the partially standardized statistics $M(k, l)$ and M,

$$\tilde{M}(k,l) = \max_{1 \le j \le p} \left(\sqrt{\frac{n_l}{n_k + n_l}} \frac{\tilde{S}_{k,j}}{\sigma^\tau_{k,l,j}} - \sqrt{\frac{n_k}{n_k + n_l}} \frac{\tilde{S}_{l,j}}{\sigma^\tau_{k,l,j}} \right) \quad \text{and}$$

$$\tilde{M} = \max_{(k,l) \in \mathcal{P}} \tilde{M}(k,l).$$

The following two theorems, with proofs provided in the supplementary material, extend the Gaussian and bootstrap approximation results in Lopes, Lin, and Müller (2020) to the multiple-sample setting as encountered in MANOVA, where d_{K} denotes the Kolmogorov distance, defined by $d_{\mathrm{K}}(\mathcal{L}(U), \mathcal{L}(V)) = \sup_{t \in \mathbb{R}} |\mathbb{P}(U \le t) - \mathbb{P}(V \le t)|$ for generic random variables U and V with probability distributions $\mathcal{L}(U)$ and $\mathcal{L}(V)$. As discussed in Section 1, this extension from the one- to the multi-sample case is nontrivial. The key theoretical results are in Theorems 3.1 and 3.2, which provide theoretical justifications for the proposed bootstrap procedure. In these theorems, the constant δ may be taken to be arbitrarily small, and so the convergence rates are nearly parametric.

Theorem 3.1 (Gaussian approximation). Fix any small $\delta > 0$, and suppose that Assumptions 1–3 hold. Then,

$$d_{\mathrm{K}}\left(\mathcal{L}(M), \mathcal{L}(\tilde{M})\right) \lesssim n^{-\frac{1}{2}+\delta}.$$

Theorem 3.2 (Bootstrap approximation). Fix any small $\delta > 0$, and suppose that Assumptions 1–3 hold. Then there is a constant $c > 0$, not depending on n, such that the event

$$d_{\mathrm{K}}\left(\mathcal{L}(\tilde{M}), \mathcal{L}(M^\star | X)\right) \le cn^{-\frac{1}{2}+\delta}$$

occurs with probability at least $1 - cn^{-1}$, where $\mathcal{L}(M^\star|X)$ represents the distribution of $M^\star$ conditional on the observed data.

3.2. High-Dimensional MANOVA

We first analyze the power of the proposed method in Section 2. All proofs are deferred to the supplementary material.

Theorem 3.3. If Assumptions 1–3 hold and the number of bootstrap samples satisfies $B \gtrsim \log^2 n$, then the following statements are true.

(i) For any fixed $\varrho \in (0, 1)$, we have $|\hat{q}_M(\varrho)| \le c \log^{1/2} n$ with probability at least $1 - cn^{-1}$, where c is a constant not depending on n.

(ii) For some constant $c > 0$ not depending on n, we have

$$\Pr\left(\max_{(k,l) \in \mathcal{P}} \max_{1 \le j \le p} \hat{\sigma}^2_{k,l,j} < 2\sigma^2_{\max} \right) \ge 1 - cn^{-1},$$

where $\sigma_{\max} = \max\{\sigma_{k,j} : 1 \le j \le p, 1 \le k \le K\}$.

Consequently, if $\max_{(k,l) \in \mathcal{P}} \max_{1 \le j \le p} |\mu_k(j) - \mu_l(j)| \ge c\sigma_{\max} n^{-1/2} \log^{1/2} n$ for a sufficiently large positive constant c not depending on n, then for any choice of $\mathcal{P}$, the null hypothesis will be rejected with probability tending to one as $n \to \infty$.

To analyze the size of the proposed test, we observe that when we construct the SCRs, we use $\hat{\sigma}_{k,l,j}$ instead of $\sigma_{k,l,j}$. This requires us to quantify the Kolmogorov distance between the distributions of M and

$$\hat{M} = \max_{(k,l) \in \mathcal{P}} \hat{M}(k,l), \tag{4}$$

where

$$\hat{M}(k,l) = \max_{1 \le j \le p} \left(\sqrt{\frac{n_l}{n_k + n_l}} \frac{S_{k,j}}{\hat{\sigma}^\tau_{k,l,j}} - \sqrt{\frac{n_k}{n_k + n_l}} \frac{S_{l,j}}{\hat{\sigma}^\tau_{k,l,j}} \right). \tag{5}$$

Note that like M defined in Equation (2), the random variable $\hat{M}$ itself is not a test statistic. With $F_{k,j}$ denoting the cumulative distribution function of the standardized random variable $\{X_{k,1}(j) - \mu_k(j)\}/\sigma_{k,j}$, we require the following mild condition on the distribution of the standardized observations.

Assumption 4. There are positive constants ν, r_0, and c not depending on n such that $\max_{1 \le k \le K} \max_{1 \le j \le p} \sup_{x \in \mathbb{R}} \sup_{r \in (0, r_0)} r^{-\nu} \Big(F_{k,j}(x + r) - F_{k,j}(x - r) \Big) \le c.$

The above condition is essentially equivalent to common Hölder continuity of the distribution functions $F_{k,j}$, that is, there is a common Hölder constant ν that is fixed but could be arbitrarily small. The assumption is satisfied if each of the distributions $F_{k,j}$ has a density function $f_{k,j}$ such that $\max_{1\le k\le K}\max_{1\le j\le p}\|f_{k,j}\|_\infty \lesssim 1$, where $\|\cdot\|_\infty$ is the supremum norm. However, the condition is much weaker than this, as it may hold even when the distributions do not have densities, or the densities are unbounded.

Theorem 3.4. Fix any small $\delta > 0$, and suppose that Assumptions 1–4 hold. Then,

$$d_{\mathrm{K}}(\mathcal{L}(\hat{M}), \mathcal{L}(M)) \lesssim n^{-\frac{1}{2}+\delta}.$$

With the triangle inequality, the above theorem together with Theorem 3.1 and 3.2 implies that, with probability at least $1 - cn^{-1}$, we have $d_{\mathrm{K}}(\mathcal{L}(\hat{M}), \mathcal{L}(M^\star \mid X)) \le cn^{-\frac{1}{2}+\delta}$, for some constant $c > 0$ not depending on n. This allows us to quantify the convergence rate of the size of the test, as follows. Let $\textsc{size}(\varrho)$ be the probability that $\mathbf{H}_0$ is rejected at the level ϱ when it is true. When $B \gtrsim n$, the Dvoretzky–Kiefer–Wolfowitz–Massart inequality (Dvoretzky, Kiefer, and Wolfowitz 1956; Massart 1990) implies that the empirical distribution of B independent samples of $M^\star$ uniformly converges to the distribution of $M^\star$ at the rate $n^{-1/2+\delta}$ with probability at least $1-cn^{-1}$. The following result is then a direct consequence of Theorems 3.1–3.4 and it asserts that the size of the test is asymptotically correctly controlled at the rate $n^{-1/2+\delta}$.

Theorem 3.5. Fix any small $\delta > 0$, and fix any $\varrho \in (0, 1)$. If Assumptions 1–4 hold, with $B \gtrsim n$, then

$$|\textsc{size}(\varrho) - \varrho| \lesssim n^{-1/2+\delta}.$$

We note that in Theorems 3.4 and 3.5, Assumption 4 can be replaced with the condition $n^{-1/2}\log^3 p \ll 1$ which then imposes an upper bound on the growth rate of p relative to n. In conjunction with the consistency of the general test as in Theorem 3.3, Theorem 3.5 provides strong justification for the application of the proposed test for a large class of null hypotheses that are typically all of interest in MANOVA in addition to the main global null hypothesis that all means are equal.

4. Application to Functional ANOVA

Consider a separable Hilbert space $\mathcal{H}$ and a second-order random element Y with mean element $\mu \in \mathcal{H}$, that is, $\mathbb{E}\|Y\|^2_{\mathcal{H}} < \infty$, where $\|\cdot\|_{\mathcal{H}}$ denotes the norm of the Hilbert space. In our context, the random element Y represents an observed functional data atom drawn from a population of functional data. Commonly considered Hilbert spaces in the area of functional data analysis include reproducing kernel Hilbert spaces and the space $L^2(\mathcal{T})$ of squared integrable functions defined on a domain $\mathcal{T}$. In one-way functional ANOVA, one aims to test the hypothesis

$$\mathbf{H}_0 : \mu_1 = \cdots = \mu_K, \qquad (6)$$

given K independent groups of iid elements $Y_{k,1}, \ldots, Y_{k,n_k} \in \mathcal{H}$ with common mean element $\mu_k \in \mathcal{H}$, with $k = 1, \ldots, K$.

Given an orthonormal basis $\phi_1, \phi_2, \ldots$ of $\mathcal{H}$, each μ_k may be represented in terms of this basis, that is, $\mu_k = \sum_{j=1}^\infty u_{kj}\phi_j$, where $u_{k,j}$ are generalized Fourier coefficients. Then the null hypothesis (6) is equivalent to the statement that $u_{k,j} = u_{l,j}$ for all $j \ge 1$ and all $1 \le k < l \le K$. This suggests that in empirical situations we choose a large integer $p \ge 1$ and test whether the vectors $u_k \equiv (u_{k,1}, \ldots, u_{k,p})$ are equal for $k = 1, \ldots, K$, which is precisely the hypothesis testing problem introduced in Section 2. This idea of transforming a functional ANOVA problem into a MANOVA problem has been proposed by Górecki and Smaga (2015) with a classic standard MANOVA method. Here we modify this idea with the proposed MANOVA method to exploit the inherited decay in variances for functional data. We first observe that each Y_k admits the Karhunen–Loève expansion $Y_k = \mu_k + \sum_{j=1}^\infty \xi_{k,j}\varphi_j$, where $\varphi_1, \varphi_2, \ldots$ are orthonormal elements of $\mathcal{H}$, and ξ_{kj} are uncorrelated random variables such that $\mathbb{E}\xi_{kj} = 0$ and $\sum_{j=1}^\infty \mathrm{var}(\xi_{kj}) < \infty$. This implies that $\mathrm{var}(\xi_{kj})$ decays to zero at a rate faster than j^{-1}. Consequently, Proposition 2.1 of Lopes, Lin, and Müller (2020) asserted that the variance of the (random) generalized Fourier coefficient of Y_k with respect to the basis element ϕ_j also decays, which allows us to adopt the test proposed in Section 2.

4.1. Simulation Studies

We assess the above method in terms of its finite sample performance by numerical simulations and compare it with three popular methods in the literature, namely the L^2 based method (L2) (Faraway 1997; Zhang and Chen 2007), the F-statistic-based method (F) (Shen and Faraway 2004; Zhang 2011) and the global pointwise F test (GPF) (Zhang and Liang 2014). These were briefly reviewed in the introduction and numerical implementations are available from Górecki and Smaga (2019), see also Górecki and Smaga (2015). We also compare it with the random projection-based method (RP) (Cuesta-Albertos and Febrero-Bande 2010), the global envelope test (GET) (Mrkvička et al. 2020) and a method (MPF) recently developed by Zhang et al. (2019a) that takes the maximum of the pointwise F-statistics as a test statistic and also leverages bootstrapping to approximate the critical value of the test.

In the simulation study, we set $\mathcal{H} = L^2([0, 1])$, and consider four families of mean functions, parameterized by $\theta \in [0, 1]$, as follows,

(M1) $\mu_k(t) = \mu_0(t) + \theta k\sum_{j=1}^{10} j^{-2}\{\sin(2j\pi t) + \cos(2j\pi t)\}/50$ with $\mu_0(t) = 5(t - 1/2)^2$,
(M2) $\mu_k(t) = \mu_0(t) + \theta k/40$ with $\mu_0(t) \equiv 1$,
(M3) $\mu_k(t) = \mu_0(t) + \theta k\{1 + (10t - 2)(10t - 5)(10t - 8)\}/40$ with $\mu_0(t) = -(f_{1/4,1/10}(t) + f_{3/4,1/10}(t))$,
(M4) $\mu_k(t) = \mu_0(t) + \theta k\exp\{-(t - 1/2)^2/100\}/25$ with $\mu_0(t) = \exp\{\sin(2\pi t)\}/2$,

for $k = 1, 2, 3$, where $f_{a,b}$ denotes the probability density function of the normal distribution with mean a and variance b^2. Obviously μ_1, μ_2, μ_3 are identical and equal to μ_0 when $\theta = 0$, and differ from each other when $\theta \ne 0$. These families are shown in Figure 2. Mean function families (M1) and (M2) represent "sparse alternatives" in the frequency domain in the

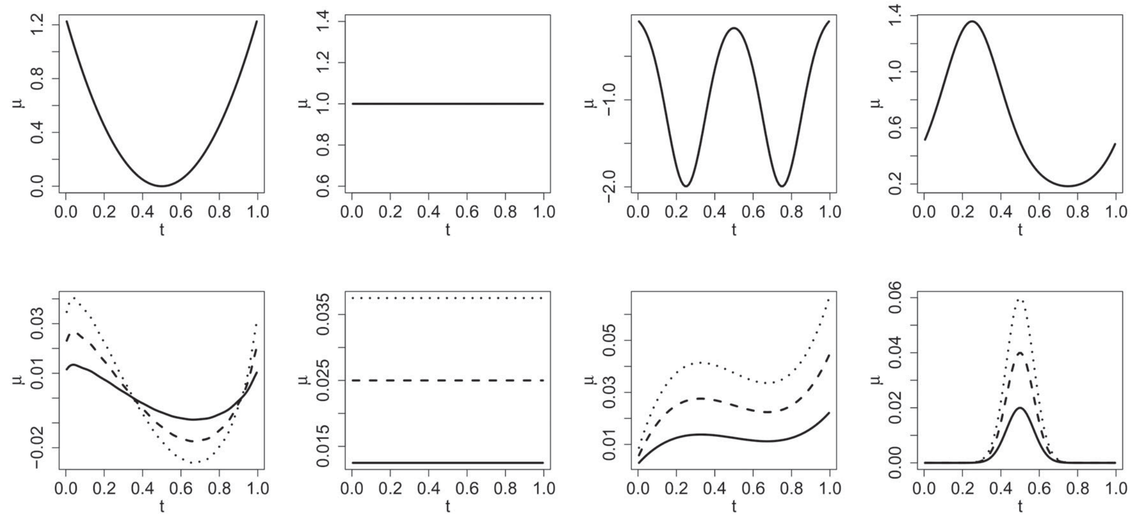

Figure 2. Mean functions. The first row shows the functions μ_0 employed for families (M1)–(M4), respectively, and the second row displays the functions $\mu_1 - \mu_0$ (solid), $\mu_2 - \mu_0$ (dashed) and $\mu_3 - \mu_0$ (dotted) with $\theta = 0.5$ in the families (M1)–(M4), respectively, from left to right.

sense that the Fourier coefficients of the mean functions differ most in the first few leading terms under the alternative when $\theta \neq 0$, while the function family (M3) represents a "dense alternative" in the frequency domain. When $\theta \neq 0$, the families (M1)–(M3) are "dense" in the time domain. In particular, the alternatives in (M2) are uniformly dense in the time domain, in the sense that the differences of the mean functions between the groups are nonzero and uniform in $t \in \mathcal{T} = [0, 1]$. Thus, families (M1)–(M3) favor the integral-based methods such as the L2, F, and GPF tests, as these methods integrate certain statistics over the time domain. In contrast, the alternatives in the last family (M4) are "sparse" in the time domain.

We sample functional data of the form $\mu_k(\cdot) + W_k(\cdot)$, for certain choices of centered random processes $W_k(\cdot)$ in two different settings. In the first "common covariance" setting, the random processes of all groups are Gaussian with the following common Matérn covariance function:

$$\mathcal{C}(s,t) = \frac{\sigma^2}{16} \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}|s-t|}{\eta} \right)^{\nu} B_\nu \left(\frac{\sqrt{2\nu}|s-t|}{\eta} \right), \qquad (7)$$

where Γ is the gamma function, B_ν is the modified Bessel function of the second kind, σ^2 is set to 2.5, η is set to 1, and ν is set to $1/2$. In the "group-specific covariance" setting, the groups have different covariance functions, as follows. For the first group, the random process is the Gaussian process with the Matérn covariance function (7). For the second group, the process is the Wiener process with dispersion $\sigma = 0.1$, that is, the Gaussian process with the covariance function $\mathcal{C}(s,t) = \sigma^2 \min(s,t)$. For the third group, we set $W_3(\cdot) = \sum_{j=1}^{51} \xi_j \phi_j(\cdot)/20$, where $\phi_1(t) \equiv 1$, $\phi_{2j} = \sin(2j\pi t)$ and $\phi_{2j+1} = \cos(2j\pi t)$, and ξ_j follows a uniform distribution on $[-j^{-2}\sqrt{3}, j^{-2}\sqrt{3}]$, providing a non-Gaussian case. All sampled functions are observed at $m = 100$ equally spaced points on the interval $[0, 1]$. Using larger values of m does not have much effect on the performance; this is in agreement with the findings in Zhang et al. (2019a).

We set the significance level at $\varrho = 0.05$, consider balanced sampling with $n_1 = n_2 = n_3 = 50$ and also unbalanced sampling with $(n_1, n_2, n_3) = (30, 50, 70)$, and use the aforementioned basis $\phi_1(t), \ldots, \phi_p(t)$ with $p = 51$. The parameter τ is selected by the method described in Section 2 from 11 candidate values, namely, $0, 0.1, \ldots, 0.9, 0.99$. Each simulation setup is replicated 1000 times independently. The results for the size of the global test are summarized in Table 1, showing that the proposed method and most of the other methods have an empirical size that is reasonably close to the nominal level. The

Table 1. Empirical size of functional ANOVA.

Covariance	M	(n_1, n_2, n_3)	Proposed	L2	F	GPF	MPF	GET	RP
Common	M1	50,50,50	0.051	0.054	0.052	0.053	0.043	0.049	0.038
		30,50,70	0.053	0.056	0.057	0.056	0.055	0.033	0.035
	M2	50,50,50	0.042	0.046	0.041	0.044	0.043	0.034	0.022
		30,50,70	0.057	0.058	0.052	0.054	0.039	0.048	0.037
	M3	50,50,50	0.057	0.056	0.050	0.054	0.047	0.036	0.023
		30,50,70	0.056	0.057	0.053	0.055	0.049	0.049	0.033
	M4	50,50,50	0.046	0.048	0.044	0.050	0.038	0.037	0.028
		30,50,70	0.053	0.054	0.052	0.051	0.045	0.041	0.028
Group-specific	M1	50,50,50	0.055	0.055	0.052	0.058	0.056	0.050	0.026
		30,50,70	0.043	0.035	0.031	0.044	0.041	0.049	0.037
	M2	50,50,50	0.056	0.059	0.056	0.061	0.057	0.054	0.034
		30,50,70	0.052	0.047	0.044	0.052	0.039	0.055	0.033
	M3	50,50,50	0.051	0.054	0.053	0.055	0.052	0.053	0.036
		30,50,70	0.049	0.043	0.039	0.048	0.045	0.066	0.030
	M4	50,50,50	0.052	0.041	0.040	0.042	0.044	0.057	0.038
		30,50,70	0.050	0.040	0.039	0.049	0.054	0.056	0.026

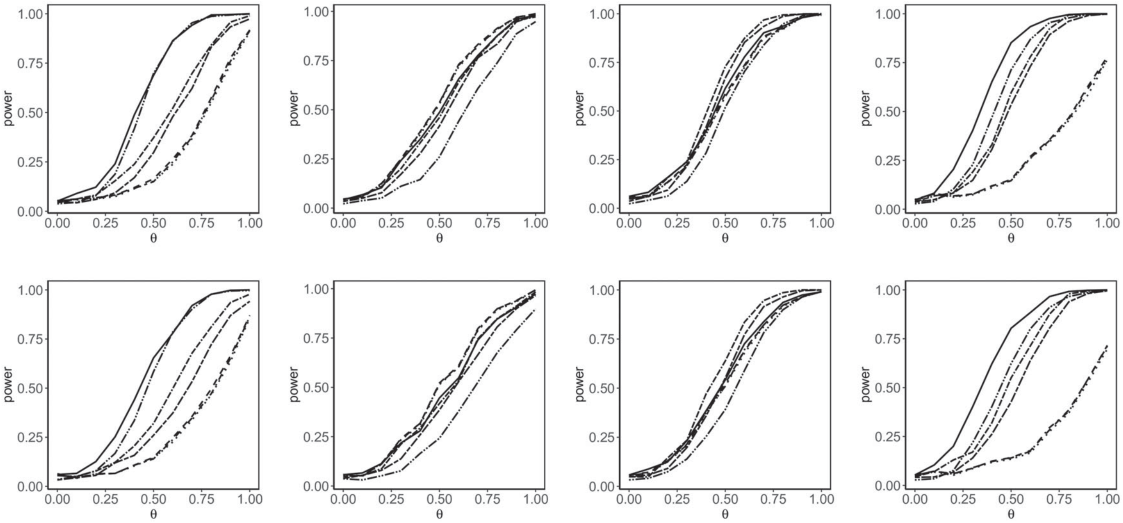

Figure 3. Empirical power of the proposed functional ANOVA (solid), L2 (dashed), F (dotted), GPF (dot-dashed), MPF (dot-dash-dashed), GET (short-long-dashed), and RP (dot-dot-dashed) in the "common covariance" setting. Top: from left to right the panels display the empirical power functions for families (M1), (M2), (M3), and (M4), when $n_1 = n_2 = n_3 = 50$. Bottom: from left to right the panels display the empirical power functions for families (M1), (M2), (M3), and (M4) for unbalanced designs when $n_1 = 30, n_2 = 50$ and $n_3 = 70$. The power functions of L2, F and GPF are nearly indistinguishable.

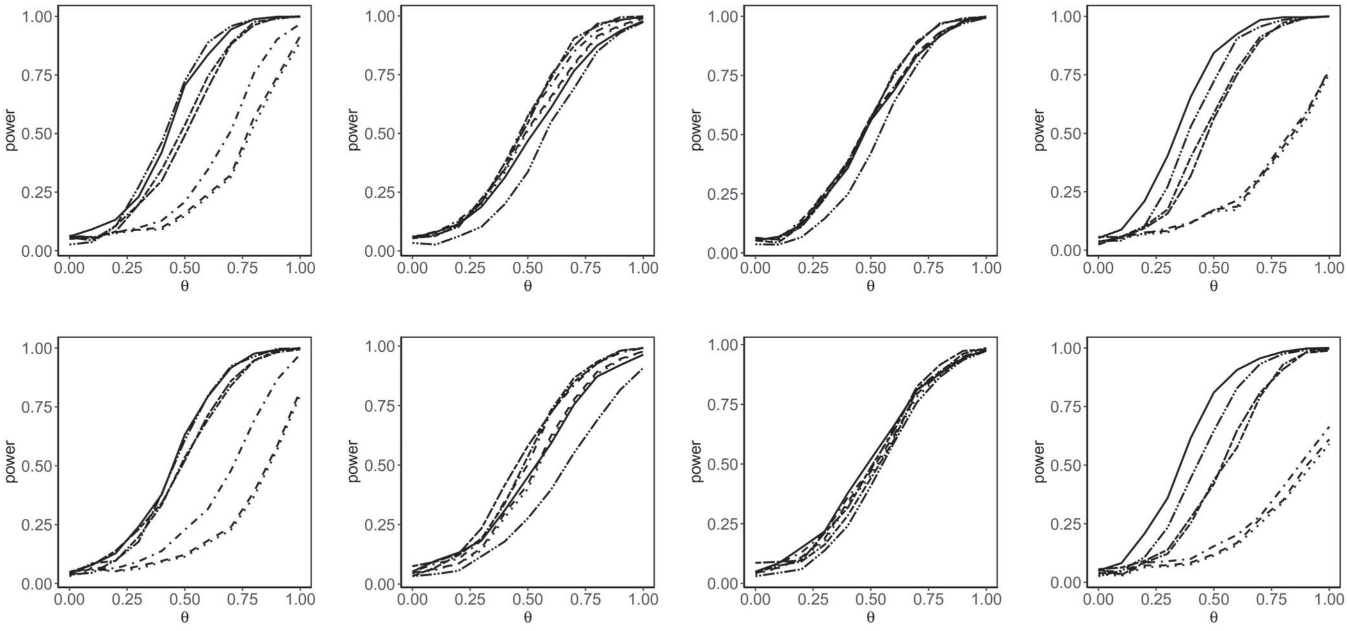

Figure 4. Same as Figure 3 but for the case of covariance functions that differ between groups.

performance in terms of power is depicted in Figure 3 for the scenario with common covariance structure. The average of the selected values for τ is 0.713 ± 0.155 and 0.754 ± 0.172 for the scenarios with common covariance structure and group-specific covariance structure, respectively.

When the alternatives are sparse in the frequency domain but not uniformly dense in the time domain (as in (M1)), or when the alternatives are sparse in the time domain (as in (M4)), the proposed method clearly outperforms most existing methods in terms of power by a large margin. The only exception is the RP method, which has similar power in the case of (M1). For the family (M2), all methods have nearly indistinguishable power, except for the RP method, which has substantially lower power. For the family (M3), the power of MPF is slightly larger in relation to the other methods. Similar observations emerge for the scenario of group-specific covariance functions with results shown in Figure 4, except that the power of GPF and MPF is slightly larger when the family is (M2), where the alternatives are uniformly dense in the time domain. In the group-specific context, the power of MPF is closer to the power of the proposed method for (M1), while the power of all methods except the RP method is nearly indistinguishable for (M3). In conclusion, the proposed test is powerful against both dense and sparse alternatives in either time or frequency domain, and provides strong improvements over existing methods in the important case where the alternative is sparse in the time domain or in the frequency domain (but not uniformly dense in the time domain).

Table 2. Computation times for functional ANOVA (in seconds).

Proposed (no GPU)	Proposed (GPU)	L2	F	GPF	MPF	GET	RP
4.792	0.085	0.002	0.002	0.005	3.602	1.629	0.877

The average computation time to complete a single Monte Carlo simulation replicate in seconds, including selecting the parameter τ by the proposed data-driven procedure in Section 2 is presented in Table 2. It shows that a single simulation replicate can be completed within 5 sec without GPU acceleration and within only 0.1 sec when using an NVIDIA Quadro P400 graphics card. Following the suggestion of a reviewer, we also investigated the impact of within-function correlation on the power by using the simulation models from Zhang et al. (2019a) and found that the proposed method is preferred when the within-function correlation is strong; see Section F of the supplementary material for details, where we also examined the effectiveness of the data-driven selection procedure for τ proposed in Section 2.

4.2. Data Application

We apply the proposed method to analyze the functional data described in Carey et al. (2008) concerning egg-laying trajectories for Mexican fruit flies (*Anastrepha ludens*) under various diets, with further perspective and background provided in Carey et al. (1998, 2002). In this study, newly merged female flies were placed in individual glass cages and during their entire lifespan were fed different diets. The number of eggs laid by each individual fly on each day was recorded and the resulting trajectories of daily egg-laying were then viewed as functional data. Since flies started egg-laying only around day 10 after emergence and to avoid selection effects due to individually varying age-at-death, we considered the trajectories on a domain [10, 50] days and included only those flies that were still alive at the right endpoint at age 50 days.

Of interest is the effect of the amount of protein in the diet on the egg-laying trajectory, as female flies require protein to produce eggs. We compare three cohorts of fruit flies which all received an overall reduced diet at 25% of full level and three different protein levels, with sugar-to-protein ratios of 3:1, 9:1, and 24:1, corresponding to fractions of 25%, 10%, and 4% of protein in the diet. The cohorts consist of $n_1 = 25, n_2 = 41$, and $n_3 = 50$ flies, respectively and are thus unbalanced. The sample mean functions for the three cohorts are depicted in Figure 5, where the noisy character of the data is reflected in the fluctuations of the functions. The mean of the cohort under a 4% protein diet is seen to be substantially smaller than the means for the other two groups, indicating that egg production is severely impeded if flies receive only 4% protein. The mean functions for the cohorts receiving 10% and 25% are much closer, indicating that protein levels above 10% have a relatively much smaller impact on egg-laying trajectories than protein levels declining below 10%.

These visual impressions are confirmed when applying the proposed functional ANOVA approach. The selected value for τ was $\tau = 0.4$ and 51 Fourier basis functions are used to represent the data. The overall p value for the null hypothesis that the three mean functions are the same is $p < 10^{-7}$ from Table 3. The pairwise comparisons between the groups with 25% protein and the 4% protein as well as between the 10% protein group and the 4% protein group show significant differences, while this is not the case for the comparison between the 25% protein group and the 10% protein group. This confirms that there is a minimum protein level that needs to be maintained as otherwise egg-laying is impeded over the entire lifespan, while more than 10% protein does not lead to major changes in the expected egg-laying trajectory. This valuable extra information is obtained without performing additional hypothesis tests and thus no requirement for adjustments for multiple comparisons that might lower the power of the test.

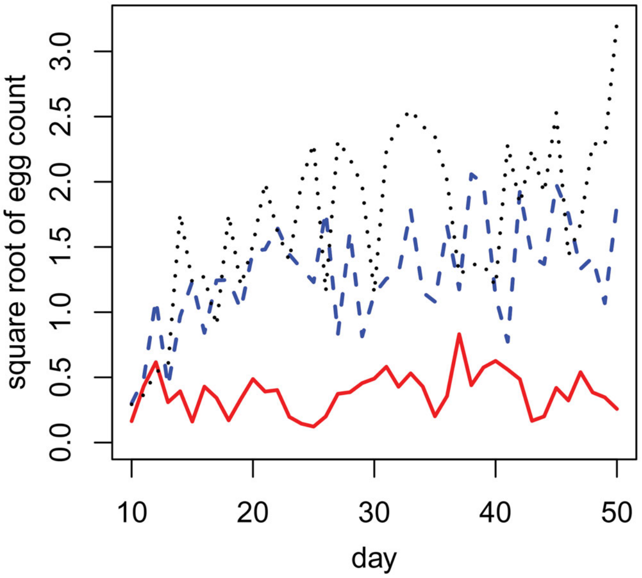

Figure 5. Sample mean trajectories of the number of eggs laid between age 10 and 50 days by female fruit flies under three different diets, where the dotted curve corresponds to a cohort of $n_1 = 25$ flies receiving a diet with 25% protein, the dashed curve to a cohort of $n_2 = 41$ flies under a diet with 10% protein, and the solid curve to a cohort of $n = 50$ flies under a diet with 4% protein.

5. Application to Sparse Count Data

Count data, often modeled by multinomial or Poisson distributions, occur in many applications. For the multinomial model, the decay in variance is an inherent feature due to the requirement that the sum of the probabilities of all categories is one. For the Poisson distribution, since the variance is equal to the mean, sparseness in the mean induces decay in the variance. Here, sparseness refers to situations where there are only a few nonzero coordinates, or where the ordered mean coordinates decrease to zero. For instance, in the field of text mining or information retrieval in which word frequency is an important feature, words in a vocabulary often have drastically different frequencies. In addition, the frequency of words decreases rapidly when moving from frequent to rare words. For example,

Table 3. *p*-values for the study on the egg-laying trajectories.

Proposed	L2	F	GPF	MPF	GET	RP
$< 10^{-7}$	3.0×10^{-15}	2.4×10^{-14}	2.4×10^{-13}	0.012	0.0005	0.0007

for the English language, the ordered word frequency is found to approximately follow Zipf's law (Zipf 1949). Below we assess the performance of the proposed method for sparse Poisson data via simulation studies and two real data applications.

5.1. Simulation Studies

We considered three groups, represented by the p-dimensional random vectors X_1, X_2, and X_3. Each random vector X_k follows a multivariate Poisson distribution (Inouye et al. 2017) and is represented by $(W_{k0} + W_{k1}, \ldots, W_{k0} + W_{kp})$, where for $k = 1, 2, 3$, $W_{k0}, \ldots, W_{kp}$ are independent Poisson random variables with mean $\eta_{k0}, \ldots, \eta_{kp} \in \mathbb{R}$, respectively. Then the jth coordinate of X_k follows also a Poisson distribution with mean $\eta_{k0} + \eta_{kj}$. In addition, all coordinates are correlated due to the shared random variable W_{k0}. In our study, we set $\eta_{k0} = 1$ for $k = 1, 2, 3$, and consider two settings for $\eta_{k1}, \ldots, \eta_{kp}$. In the first "sparse" setting, $\eta_{kj} = (1 + \theta k)j^{-1}$ for $k = 1, 2, 3$ and $j = 1, \ldots, p$. In this setting, when $\theta \neq 0$, the difference of the mean in the jth coordinate decays as j^{-1}. In the second "dense" setting, we set $\eta_{kj} = j^{-1} + \theta k/2$, so that the difference of the mean in each coordinate is equal. Note that the setting with $\theta = 0$ corresponds to the null hypothesis, under which the mean vectors of all groups are identical. For the dimension, we consider two cases, namely $p = 25$ and $p = 100$, and for sample size the balanced case $(n_1, n_2, n_3) = (50, 50, 50)$ and an unbalanced case with $(n_1, n_2, n_3) = (30, 50, 70)$. The parameter τ is selected by the method described in Section 2. Each simulation is repeated 1000 times. Across all settings, the average value of selected τ is 0.305 ± 0.221 and 0.341 ± 0.237 for $p = 25$ and $p = 100$, respectively.

For comparison purposes, we implemented the procedure (S) of Schott (2007) and the data-adaptive ℓ_p-norm-based test (DALp) (Zhang et al. 2018) that are reviewed in the introduction. The former is based on the limit distribution of a test statistic that is composed of inter-group and within-group sums of squares, while the latter utilizes an adjusted ℓ_p-norm-based test statistic whose distribution is approximated by a multiplier bootstrap. The former is favored for testing problems with a dense alternative, while the latter has been reported to be powerful against different patterns of alternatives (Zhang et al. 2018). We also include the classic Lawley–Hotelling trace test (LH) (Lawley 1938; Hotelling 1947) as a baseline method which is not specifically designed for the high-dimensional setting, and its ridge-regularized version (RRLH) (Li, Aue, and Paul 2020) targeting the high-dimensional scenario. The empirical sizes in Table 4 demonstrate that those of the proposed test and the test of Schott (2007) are quite close to the nominal level, while the size of the test of Zhang et al. (2018) seemed slightly inflated and the sizes of the Lawley–Hotelling trace test and its regularized version are rather conservative in the high-dimensional case $p = 100$. The power function for the sparse case $(n_1, n_2, n_3) = (30, 50, 70)$ is shown in Figure 6, while the power function for $(n_1, n_2, n_3) = (50, 50, 50)$ is very similar (not shown). One finds that in the sparse case, the proposed test has substantially more power than the test of Zhang et al. (2018), while the latter in turn has more power than the test of Schott (2007) and the Lawley–Hotelling trace tests. In the dense setting which does not favor the proposed test, it is seen to have power behavior that is comparable with that of the tests of Schott (2007) and Zhang et al. (2018), and all of these methods outperform the Lawley–Hotelling trace test whose performance substantially deteriorates for higher dimensions. The regularized Lawley–Hotelling trace test substantially improves upon the classic version only in the sparse setting and when the dimension is relatively large, for example, when $p = 100$. The average computation time to complete a single Monte Carlo simulation replicate is presented in Table 5, where $p = 100$ and the parameter τ is selected from 11 candidate values by the data-driven procedure proposed in Section 2. We observe that a single simulation replicate can be completed within 10 sec without GPU acceleration and within 0.2 sec by utilizing an NVIDIA Quadro P400 graphics card. In addition to testing hypotheses, the proposed method can also simultaneously identify the pairs of groups, as well as coordinates, that have significantly different means, as we demonstrate below for two real datasets.

Table 4. Empirical size of ANOVA on the Poisson data.

	p	n	Proposed	S	DALp	LH	RRLH
Sparse	25	50,50,50	0.055	0.042	0.065	0.045	0.051
		30,50,70	0.052	0.053	0.069	0.048	0.053
	100	50,50,50	0.056	0.045	0.054	0.000	0.000
		30,50,70	0.056	0.055	0.065	0.000	0.002
Dense	25	50,50,50	0.050	0.051	0.065	0.045	0.065
		30,50,70	0.045	0.066	0.062	0.050	0.050
	100	50,50,50	0.057	0.054	0.064	0.001	0.004
		30,50,70	0.051	0.049	0.067	0.001	0.000

Table 5. Average computation time for ANOVA on the Poisson data (in seconds).

Proposed (no GPU)	proposed (GPU)	S	DALp	LH	RRLH
9.869	0.155	0.011	0.461	0.030	0.135

5.2. Data Applications

We apply the proposed method to analyze the CLASSIC3 dataset[3] (Dhillon, Mallela, and Modha 2003) that has been studied in information retrieval. The data consist of 3891 document abstracts from three different domains, specifically, $n_1 = 1460$ from information retrieval (CISI), $n_2 = 1398$ from aeronautical systems (CRAN) and $n_3 = 1033$ from medical research (MED). Standard text preprocessing was applied to

[3] Originally available from ftp://ftp.cs.cornell.edu/pub/smart, and now available publicly on the Internet, for example, *https://www.dataminingresearch.com/index.php/2010/09/classic3-classic4-datasets/*

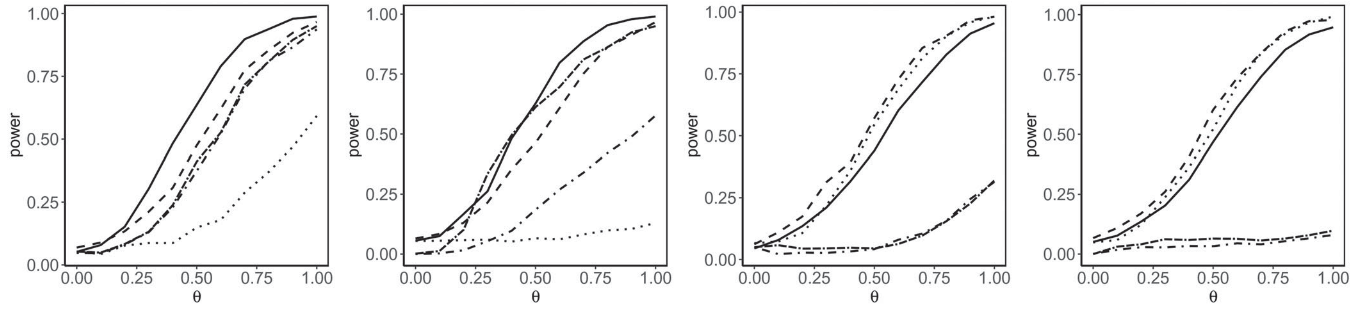

Figure 6. Empirical power of the proposed high-dimensional ANOVA (solid), DALp (dashed), S (dotted), LH (dot-dashed), and RRLH (dot-dash-dashed), when $(n_1, n_2, n_3) = (30, 50, 70)$, for the sparse setting with $p = 25$ (first panel) and $p = 100$ (second panel) and for the dense setting with $p = 25$ (third panel) and $p = 100$ (fourth panel).

Table 6. The average frequency of words that are significantly different among all categories.

	Use	Data	Pressure	Effect	Theory	Problem	Body	Increase	Normal	Group
CISI	**0.715**	**0.401**	0.011	0.060	0.167	0.301	0.017	0.089	0.007	0.129
CRAN	0.515	0.239	**1.004**	**0.759**	**0.684**	**0.456**	**0.607**	0.271	0.112	0.011
MED	0.265	0.082	0.139	0.338	0.024	0.069	0.162	**0.437**	**0.351**	**0.304**

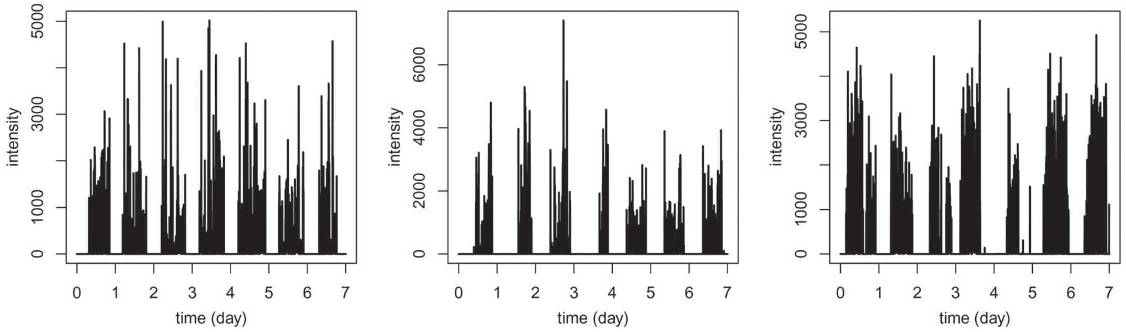

Figure 7. Activity intensity trajectories of three randomly selected participants from the NHANES data 2005–2006.

these abstracts, including removal of high-frequency common words (commonly referred to as stop words, such as "the," "is," "and," etc), punctuation and Arabic numbers. In addition, we follow common practice in the field of information retrieval to reduce inflected words to their word stem, base or root form by using a stemmer, such as the Krovetz stemmer (Krovetz 1993). Each document is then represented by a vector of word counts. These vectors are naturally sparse, as the number of distinct words appearing in a document is in general far less than the size of the vocabulary. Intuitively, vocabularies from different domains are different. Our goal is to examine this intuition and to find the words that are substantially different among the three domains. To this end, we focus on words with at least 50 occurrences in total to eliminate the effects of rare words. This results in $p = 1296$ distinct words under consideration. Then, we applied the proposed test to the processed data and found that the vocabularies used in these three domains are not the same among any pair of the domains, with p-value less than 10^{-7} where τ was selected as $\tau = 0.6$. In particular, the proposed method simultaneously identifies the words that have significantly different frequencies among the domains, which are shown in Table 6, where the numbers represent the average frequency of the words within each domain. The results for CISI and CRAN match our intuition about these two domains. For the domain of medical research, the word "normal" is often used to refer to healthy patients or subjects, while the word "increase" is used to describe the change of certain health metrics, such as blood pressure.

Next, we apply the proposed method to study physical activity using data collected by wearable devices, as available in the National Health and Nutrition Examination Survey (NHANES) 2005–2006. In the survey, each participant of age 6 years or above was asked to wear a physical activity monitor (Actigraph 7164) for seven consecutive days, with bedtime excluded. Also, as the device is not waterproof, participants were advised to remove it during swimming or bathing. The monitor detected and recorded the magnitude of acceleration of movement of the participant. For each minute, the readings were summarized to yield one single integer in the interval $[0, 32, 767]$ that signifies the average intensity of movement within that minute. This results in $m = 60 \times 24 \times 7 = 10{,}080$ observations per participant. Demographic characteristics of the participants are also available, and in our analysis we focused on two age groups and two marital categories. The two age groups are young adulthood with age ranging from 18 to 44, and middle-age adulthood with age ranging from 45 to 65. The two marital groups are "single" (including the widowed, divorced, separated and never-married categories in the original data) and "non-single" (including married and living-with-partner categories). These groups induce four cohorts: young nonsingle adults,

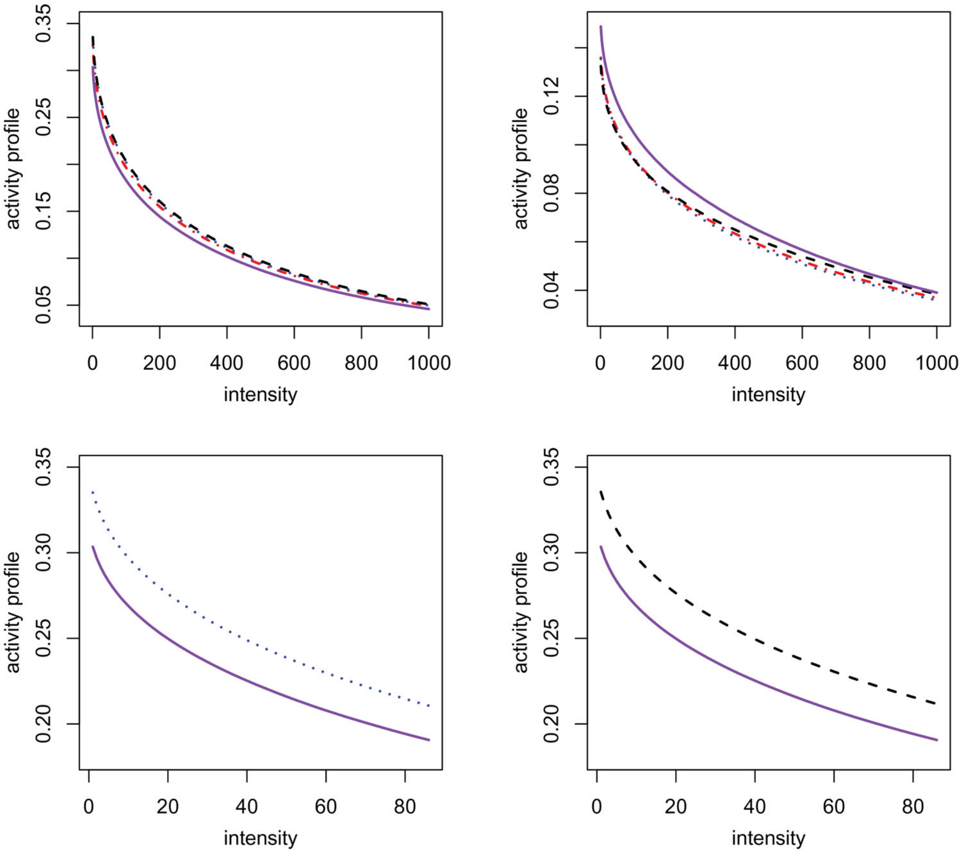

Figure 8. Top: the coordinate-wise mean activity (left) and its standard deviation (right) of young non-single cohort (dash-dotted), young single cohort (dotted), middle-age nonsingle cohort (dashed), and middle-age single cohort (solid); bottom-left: mean activity profiles of the young single cohort (dotted) and the middle-age single cohort (solid) shown for the intensity spectrum on which the differences in the means are significant among the two cohorts; bottom-right: mean activity profiles of the middle-age nonsingle cohort (dashed) and the middle-age single cohort (solid) over the spectrum on which the differences in the means are significant among the two cohorts.

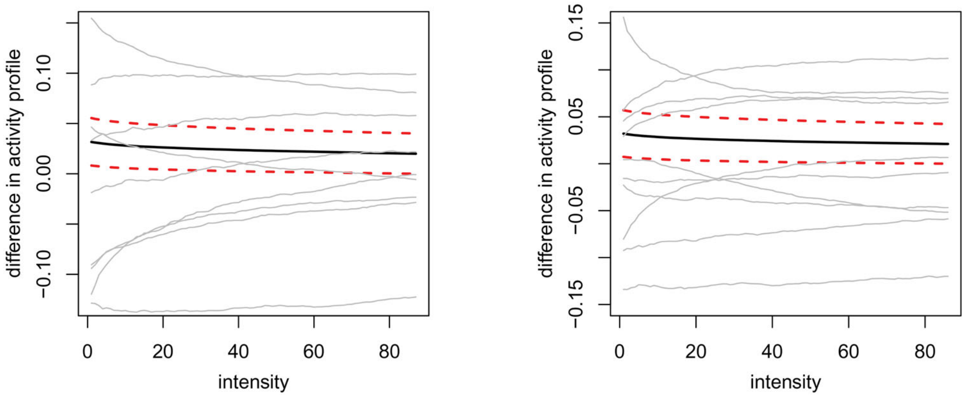

Figure 9. The empirical SCRs (dashed) for the difference (solid) of mean activity profiles over [1, 87]. The left panel corresponds to young single and middle-age single cohorts. The right panel corresponds to middle-age nonsingle and middle-age single cohorts. The light gray solid lines are differences of activity profiles of some pairs of participants from the corresponding pairs of cohorts, included to illustrate the variability of the differences in the individual level.

young single adults, middle-age non-single adults and middle-age single adults. Our goal is to examine whether the physical activity patterns are different among these cohorts.

Figure 7 presents the activity trajectories of three randomly selected participants, showing that they have different circadian rhythms. To address this problem, we adopt the strategy proposed by Chang and McKeague (2020), who studied physical activity of elder veterans from the perspective of functional data analysis, by transforming each activity trajectory $A(t)$ into an activity profile $X(j) = \mathrm{Leb}(\{t \in [0, 7] : A(t) \geq j\})$ for $j = 1, \ldots, 32{,}767$, where Leb denotes the Lebesgue measure on $\mathbb{R}$. This is essentially equivalent to accumulated $F_A(j)/m$, where $F_A(j)$ denotes the frequency of j, that is, the number of occurrences of the intensity value j, in the trajectory A. Therefore, the activity profile $X(j)$ can be viewed as count data normalized by m. As over 95% of the physical activity has low to moderate intensity, that is, with intensity value below 1000, we focus on the intensity spectrum $[1, 1000]$. In addition, we exclude subjects

Table 7. *p*-values for studies on CLASSIC3 and NHANES datasets.

	Proposed	S	DALp	LH	RRLH
CLASSIC3	$< 10^{-7}$	$0^{\dagger}$	$< 10^{-7}$	$0^{\dagger}$	$0^{\dagger}$
NHANES	0.004	0.005	0.005	0.936	0.716

†The *p*-values are below machine precision.

with readings that are missing, unreliable or from a monitor not in calibration. This results in four cohorts of size $n_1 = 1027$, $n_2 = 891$, $n_3 = 610$, and $n_4 = 339$, respectively.

The mean activity profiles and their standard deviations are depicted in the top panels of Figure 8, from which we observe that both the mean and standard deviation decay quite fast. In addition, the mean profiles from the young single and middle-age non-single cohorts are almost indistinguishable in the plot, while the mean profile of the middle-age single cohort is visibly different from the others. These visual impressions are in line with the results obtained with the proposed test, which rejects the global null hypothesis with an approximate p-value of 0.004 and thus suggests that some mean activity profiles are likely to be substantially different, where the selected value for τ is 0.5. The methods of Schott (2007) and Zhang et al. (2018) also rejected the null hypothesis with a similar p-value, while both Lawley–Hotelling trace test and its regularized version do not; see Table 7 for the detailed p-values of these methods. The proposed method also identifies two pairs of cohorts whose mean activity profiles are different and the intensity spectrum on which the differences are significant, namely, the young single cohort and the middle-age single cohort on the spectrum $[1, 87]$, and the middle-age non-single cohort and middle-age single cohort on the spectrum $[1, 86]$. These findings are visualized in the bottom panels of Figure 8. Furthermore, the proposed method provides SCRs for the differences of mean activity profiles among all pairs of cohorts. For instance, in Figure 9 we present the 95% SCRs for the pairs with differences in the mean activity profiles over the spectrum on which the differences are statistically significant. In summary, comparing to the young single and middle-age non-single cohorts, the middle-age single cohort is found to have less activity on average in the low-intensity activity spectrum.

6. Concluding Remarks

The proposed method for high-dimensional ANOVA via bootstrapping max statistics leads to the construction of SCRs for the differences of population mean vectors and is applicable for various statistical frameworks, including functional data analysis and multinomial and count data settings. The theoretical justifications rely on two key ingredients, variance decay and partial standardization, which imply near-parametric rates of convergence in high dimensions. In simulations, the resulting tests are shown to be highly competitive in terms of controlling the size of the tests and power in a variety of scenarios. It is notable that the proposed method can be completely parallelized which leads to very fast implementations on parallel processors.

As predicted by theory, performance of the proposed method is geared toward the case of sparse signals and in such scenarios it routinely outperforms competing methods in simulations. It is also found to be competitive for situations with dense signals. Since it is often unknown whether signals are sparse or dense in practice, this makes the method quite appealing for high-dimensional ANOVA in the presence of variance decay, notably for functional ANOVA problems where such variance decay is an inherent feature.

The proposed method employs a parameter τ that controls the partial standardization, which is chosen data adaptively. The implementation can be further accelerated by choosing a fixed value, where the choice $\tau = 0.8$ was shown to be effective in simulation studies in Sections F and G of the supplementary material. The principle of partial standardization may be of broader interest.

Supplementary Material

Supplement: The Supplement contains the proofs for the results in Section 3, and additional simulation studies for functional ANOVA and high-dimensional MANOVA. (PDF)

R-package: The `hdanova.cuda` package[4] implements the proposed method for the GPU based computing platform.

Funding

Research partially supported by NUS start-up grant R-155-001-217-133. Research supported in part by NSF grant DMS-1915786. Research supported in part by NSF grant DMS-2014626.

ORCID

Zhenhua Lin http://orcid.org/0000-0003-1690-9713

References

Aneiros, G., Cao, R., Fraiman, R., Genest, C., and Vieu, P. (2019), "Recent Advances in Functional Data Analysis and High-dimensional Statistics," *Journal of Multivariate Analysis*, 170, 3–9. [2]

Bai, Z., Choi, K. P., and Fujikoshi, Y. (2018), "Limiting Behavior of Eigenvalues in High-dimensional MANOVA via RMT," *The Annals of Statistics*, 46, 2985–3013. [1]

Bai, Z., and Saranadasa, H. (1996), "Effect of High Dimension: By an Example of a Two Sample Problem," *Statistica Sinica*, 6, 311–329. [1]

Cai, T. T., Liu, W., and Xia, Y. (2014), "Two-sample Test of High Dimensional Means Under Dependence," *Journal of Royal Statistical Society*, Series B, 76, 349–372. [1]

Cai, T. T., and Xia, Y. (2014), "High-dimensional Sparse MANOVA," *Journal of Multivariate Analysis*, 131, 174–196. [1]

Carey, J., Harshman, L., Liedo, P., Müller, H.-G., Wang, J.-L., and Zhen, Z. (2008), "Longevity-fertility Trade-offs in the Tephritid Fruit Fly, Anastrepha ludens, Across Dietary-restriction Gradients." *Aging Cell*, 7, 470–477. [9]

Carey, J. R., Liedo, P., Harshman, L., Zhang, Y., Müller, H.-G., Partridge, L., and Wang, J.-L. (2002), "Life History Response of Mediterranean Fruit Flies to Dietary Restriction," *Aging Cell*, 1, 140–148. [9]

Carey, J. R., Liedo, P., Müller, H.-G., Wang, J.-L., and Vaupel, J. W. (1998), "Dual Modes of Aging in Mediterranean Fruit Fly Females," *Science*, 281, 996–998. [9]

Chang, H.-W., and McKeague, I. W. (2020), "Nonparametric Comparisons of Activity Profiles From Wearable Device Data," preprint. [2,12]

[4] https://github.com/linulysses/hdanova.cuda

Chang, J., Zheng, C., Zhou, W.-X., and Zhou, W. (2017), "Simulation-based Hypothesis Testing of High Dimensional Means Under Covariance Heterogeneity," *Biometrics*, 73, 1300–1310. [1,2]

Chen, S. X., and Qin, Y.-L. (2010), "A Two-sample Test for High-dimensional Data With Applications to Gene-set Testing," *The Annals of Statistics*, 38, 808–835. [1]

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013), "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-dimensional Random Vectors," *The Annals of Statistics*, 41, 2786–2819. [1]

——— (2017), "Central Limit Theorems and Bootstrap in High Dimensions," *The Annals of Probability*, 45, 2309–2352. [1]

Cuesta-Albertos, J. A., and Febrero-Bande, M. (2010), "A Simple Multiway ANOVA for Functional Data," *Test*, 19, 537–557. [2,6]

Dhillon, I., Mallela, S., and Modha, D. (2003), "Information-Theoretic Co-clustering," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 89–98. [10]

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956), "Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator," *Annals of Mathematical Statistics*, 27, 642–669. [6]

Fan, J., and Lin, S.-K. (1998), "Test of Significance When Data Are Curves," *Journal of the American Statistical Association*, 93, 1007–1021. [2]

Faraway, J. J. (1997), "Regression Analysis for a Functional Response," *Technometrics*, 39, 254–261. [2,6]

Feng, L., and Sun, F. (2015), "A Note on High-dimensional Two-sample Test," *Statistics and Probability Letters*, 105, 29–36. [1]

Feng, L., Zou, C., Wang, Z., and Zhu, L. (2015), "Two-sample Behrens-Fisher Problem for High-dimensional Data," *Statistica Sinica*, 25, 1297–1312. [1]

Ferraty, F., and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, New York: Springer-Verlag. [2]

Fujikoshi, Y., Himeno, T., and Wakaki, H. (2004), "Asymptotic Results of a High Dimensional MANOVA Test and Power Comparison When the Dimension is Large Compared to the Sample Size," *Journal of the Japan Statistical Society*, 34, 19–26. [1]

Górecki, T., and Smaga, L. (2015), "A Comparison of Tests for the One-way ANOVA Problem for Functional Data," *Computational Statistics*, 30, 987–1010. [6]

——— (2019), "fdANOVA: An R Software Package for Analysis of Variance for Univariate and Multivariate Functional Data," *Computational Statistics*, 34, 571–597. [6]

Gregory, K. B., Carroll, R. J., Baladandayuthapani, V., and Lahiri, S. N. (2015), "A Two-Sample Test for Equality of Means in High Dimension," *Journal of the American Statistical Association*, 110, 837–849. [1]

Horváth, L., and Kokoszka, P. (2012), *Inference for Functional Data with Applications*, Springer Series in Statistics, New York: Springer. [2]

Hotelling, H. (1947), "Multivariate Quality Control Illustrated by Air Testing of Sample Bombsights," in *Techniques of Statistical Analysis*, eds. C. Eisenhart, M. W. Hastay, and W. A. Wallis, New York: McGraw-Hill, pp. 111–184. [10]

Hsing, T., and Eubank, R. (2015), *Theoretical Foundations of Functional Data Analysis, With an Introduction to Linear Operators*, Chichester: Wiley. [2]

Hu, J., Bai, Z., Wang, C., and Wang, W. (2017), "On Testing the Equality of High Dimensional Mean Vectors With Unequal Covariance Matrices," *Annals of the Institute of Statistical Mathematics*, 69, 365–387. [1]

Inouye, D. I., Yang, E., Allen, G. I., and Ravikumar, P. (2017), "A Review of Multivariate Distributions for Count Data Derived From the Poisson Distribution," *WIREs Computational Statistics*, 9, e1398. [10]

Kokoszka, P., and Reimherr, M. (2017), *Introduction to Functional Data Analysis*, Boca Raton: Chapman and Hall/CRC. [2]

Krovetz, R. (1993), "Viewing Morphology as an Inference Process," in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York: ACM Press, pp. 191–202. [11]

Lawley, D. N. (1938), "A Generalization of Fisher's z Test," *Biometrika*, 30, 180–187. [10]

Li, H., Aue, A., and Paul, D. (2020), "High-dimensional General Linear Hypothesis Tests Via Non-linear Spectral Shrinkage," *Bernoulli*, 26, 2541–2571. [1,10]

Li, H., Hu, J., Bai, Z., Yin, Y., and Zou, K. (2017), "Test on the Linear Combinations of Mean Vectors in High-dimensional Data," *Test*, 26, 188–208. [1]

Lopes, M. E., Jacob, L., and Wainwright, M. J. (2011), "A More Powerful Two-sample Test in High Dimensions Using Random Projection," in *Advances in Neural Information Processing Systems*, pp. 1206–1214. [1]

Lopes, M. E., Lin, Z., and Müller, H.-G. (2020), "Bootstrapping Max Statistics in High Dimensions: Near-parametric Rates Under Weak Variance Decay and Application to Functional Data Analysis," *The Annals of Statistics*, 48, 1214–1229. [1,2,3,5,6]

Massart, P. (1990), "The Tight Constant in the Dvoretzky–Kiefer–Wolfowitz Inequality," *Annals of Probability*, 18, 1269–1283. [6]

Mrkvička, T., Myllymäki, M., Jílek, M., and Hahn, U. (2020), "A One-way ANOVA Test for Functional Data With Graphical Interpretation," *Kybernetika*, 56, 432–458. [2,6]

Paparoditis, E. and Sapatinas, T. (2016), "Bootstrap-based Testing of Equality of Mean Functions or Equality of Covariance Operators for Functional Data," *Biometrika*, 103, 727–733. [2]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis* (2nd ed.), Springer Series in Statistics, New York: Springer. [2]

Schott, J. R. (2007), "Some High-dimensional Tests for a One-way MANOVA," *Journal of Multivariate Analysis*, 98, 1825–1839. [1,10,13]

Shen, Q., and Faraway, J. (2004), "An F Test for Linear Models With Functional Responses," *Statistica Sinica*, 14, 1239–1257. [2,6]

Srivastava, M. S., and Fujikoshi, Y. (2006), "Multivariate Analysis of Variance With Fewer Observations Than the Dimension," *Journal of Multivariate Analysis*, 97, 1927 – 1940. [1]

Srivastava, M. S., and Kubokawa, T. (2013), "Tests for Multivariate Analysis of Variance in High Dimension Under Non-Normality," *Journal of Multivariate Analysis*, 115, 204–216. [1]

Städler, N., and Mukherjee, S. (2016), "Two-sample Testing in High Dimensions," *Journal of Royal Statistical Society*, Series B, 79, 225–246. [1]

Thulin, M. (2014), "A High-dimensional Two-sample Test for the Mean Using Random Subspaces," *Computational Statistics and Data Analysis*, 74, 26–38. [1]

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016), "Functional Data Analysis," *Annual Review of Statistics and Its Application*, 3, 257–295. [2]

Xu, G., Lin, L., Wei, P., and Pan, W. (2016), "An Adaptive Two-sample Test for Highdimensional Means," *Biometrika*, 103, 609–624. [1]

Xue, K., and Yao, F. (2020), "Distribution and Correlation Free Two-sample Test of High-dimensional Means," *The Annals of Statistics*. [1,2]

Yamada, T., and Himeno, T. (2015), "Testing Homogeneity of Mean Vectors Under Heteroscedasticity in High-dimension," *Journal of Multivariate Analysis*, 139, 7 – 27. [1]

Yamada, T., and Srivastava, M. S. (2012), "A Test for Multivariate Analysis of Variance in High Dimension," *Communications in Statistics - Theory and Methods*, 41, 2602–2615. [1]

Zhang, J., and Pan, M. (2016), "A High-dimension Two-sample Test for the Mean Using Cluster Subspaces," *Computational Statistics & Data Analysis*, 97, 87 – 97. [1]

Zhang, J.-T. (2011), "Statistical Inferences for Linear Models With Functional Responses," *Statistica Sinica*, 21, 1431–1451. [2,6]

——— (2013), *Analysis of Variance for Functional Data*, London: Chapman & Hall. [2]

Zhang, J.-T., and Chen, J. (2007), "Statistical Inferences for Functional Data," *The Annals of Statistics*, 35, 1052–1079. [2,6]

Zhang, J.-T., Cheng, M.-Y., Wu, H.-T., and Zhou, B. (2019a), "A New Test for Functional One-way ANOVA With Applications to Ischemic Heart Screening," *Computational Statistics & Data Analysis*, 132, 3–17. [2,6,7,9]

Zhang, J.-T., Guo, J., and Zhou, B. (2017), "Linear Hypothesis Testing in High-dimensional One-way MANOVA," *Journal of Multivariate Analysis*, 155, 200 – 216. [1]

Zhang, J.-T., Guo, J., Zhou, B., and Cheng, M.-Y. (2019b), "A Simple Two-Sample Test in High Dimensions Based on L^2-Norm," *Journal of the American Statistical Association*. [1]

Zhang, J.-T., and Liang, X. (2014), "One-Way ANOVA for Functional Data via Globalizing the Pointwise F-test," *Scandinavian Journal of Statistics*, 41, 51–71. [2,6]

Zhang, J.-T., and Xu, J. (2009), "On the k-sample Behrens-Fisher Problem for High-dimensional Data," *Science in China, Series A: Mathematics*, 52, 1285–1304. [1]

Zhang, M., Zhou, C., He, Y., and Liu, B. (2018), "Data-adaptive Test for High-dimensional Multivariate Analysis of Variance Problem," *Australian & New Zealand Journal of Statistics*, 60, 447–470. [1,2,10,13]

Zhou, B., Guo, J., and Zhang, J.-T. (2017), "High-dimensional General Linear Hypothesis Testing Under Heteroscedasticity," *Journal of Statistical Planning and Inference*, 188, 36–54. [1]

Zipf, G. K. (1949), *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Cambridge, MA: Addison-Wesley Press. [10]