The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1292–1312
https://doi.org/10.1214/21-AOAS1537

B-SCALING: A NOVEL NONPARAMETRIC DATA FUSION METHOD

BY YIWEN LIU [1,a], XIAOXIAO SUN [1,b], WENXUAN ZHONG[2,c] AND BING LI[3,d]

[1]*Department of Epidemiology and Biostatistics, University of Arizona,* [a]*yiwenliu@arizona.edu,* [b]*xiaosun@arizona.edu*
[2]*Department of Statistics, University of Georgia,* [c]*wenxuan@uga.edu*
[3]*Department of Statistics, Pennsylvania State University,* [d]*bing@stat.psu.edu*

Very often for the same scientific question, there may exist different techniques or experiments that measure the same numerical quantity. Historically, various methods have been developed to exploit the information within each type of data independently. However, statistical data fusion methods that could effectively integrate multisource data under a unified framework are lacking. In this paper we propose a novel data fusion method, called B-scaling, for integrating multisource data. Consider K measurements that are generated from different sources but measure the same latent variable through some linear or nonlinear ways. We seek to find a representation of the latent variable, named B-mean, which captures the common information contained in the K measurements while taking into account the nonlinear mappings between them and the latent variable. We also establish the asymptotic property of the B-mean and apply the proposed method to integrate multiple histone modifications and DNA methylation levels for characterizing epigenomic landscape. Both numerical and empirical studies show that B-scaling is a powerful data fusion method with broad applications.

1. Introduction. Both the amount and variety of data have been increasing dramatically in recent years from all fields of science, such as genomics, chemistry, geophysics, and engineering. Even for the same scientific question, there may exist different techniques or experiments that measure the same numerical quantity. Historically, various methods have been developed to exploit the information within each type of data independently. However, the development of methods that could effectively integrate multisource data under a unified framework is lagging behind (Ritchie et al. (2015)). In contrast, such multisource data are abundant in practice. One motivating example is the Roadmap Epigenomics Project (Bernstein et al. (2010)).

MOTIVATING EXAMPLE (Epigenetic data from the Roadmap Epigenomics Project). As an emerging science, epigenomics is the study of the complete set of epigenetic modifications that regulate gene expressions (Bird (2007), Egger et al. (2004), Esteller (2008)). With advanced sequencing technology, the Roadmap Epigenomics Project has provided a large amount of epigenomic data of different types and structures to characterize the epigenomic landscapes of primary human tissues and cells under different conditions or diseases (Amin et al. (2015), Kundaje et al. (2015), Romanoski et al. (2015)). Chromatin immunoprecipitation (ChIP) and bisulfite treatment, for example, are two different techniques. They are adopted to generate datasets of various types of histone modification levels and DNA methylation levels (Figure 1), all of which are indicators of epigenetic activeness. Despite the availability of large amounts of data, there is still a lack of systematic understanding of how epigenomic variations across genome relate to gene expression changes (Gomez-Cabrero et al. (2014), Hamid et al. (2009), Ritchie et al. (2015)). To bridge this gap is the direct motivation of our data fusion method. We illustrated in Figure 1 the application of data fusion

Received May 2020; revised August 2021.
Key words and phrases. Data fusion, multisource data, generalized eigenvalue problem, epigenetics.

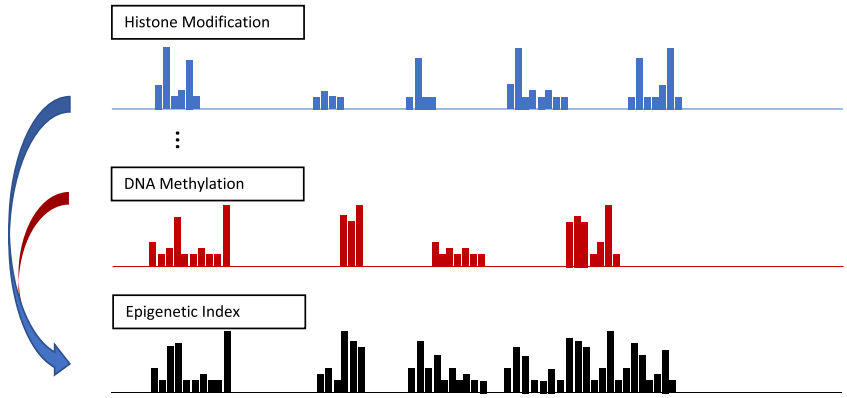

FIG. 1. *Illustration of data fusion for epigenetic data from the Roadmap Epigenomics Project. As epigenetic indicators, multiple histone modifications and DNA methylation levels (represented by the heights of the bars) across the genome are measured from the experiments. To gain a systematic understanding of the epigenetic landscape, we aim to fuse those measurements and construct an epigenetic index. Such an index is critical for understanding the epigenetic regulation of gene expression systematically.*

for epigenetic data. The goal is to fuse multiple histone modifications and DNA methylation levels and develop an epigenetic index for characterizing epigenetic landscape.

Data fusion can have numerous meanings. In this paper we refer to the process of integrating information from multiple data sources as data fusion (Hall and Llinas (1997), Hall and McMullen (2004), Khaleghi et al. (2013), Waltz, Llinas et al. (1990)). More specifically, we focus on fusions at the data level. The primary motivation of data level fusion is to provide key factors that would improve the interpretations and predictions of potential outcomes. Researchers may be able to capture some pieces of the puzzle by a single-data-type study, but data fusion approaches could help fit all pieces together. In the above example, epigenetic modifications, such as histone modification and DNA methylation, jointly characterize epigenetic variations across cell types and genome, especially in the promoter region immediately upstream of where a gene's transcription begins. Integrative analysis of these epigenetic features would considerably improve the power of using epigenetic variation to explain gene expressions. In addition, data fusion approaches could provide more efficient ways of studying the statistical associations than an analysis that uses only a single data type. Finally, since it is computationally intensive to deal with multiple large-scale datasets simultaneously, the integrative analysis that allows data integration can help to ease computation intensity while, at the same time, it preserves as much useful information of the original data as possible.

The integration of multiple data sources has been considered previously in the genomic literature. A comprehensive review of this topic can be found in Ritchie et al. (2015) and Ray, Liu and Fenyö (2017). The development has taken several paths. One may regard data integration as a dimension reduction problem and try to find a lower dimensional representation of the multiple measurements. Principal component analysis (PCA), for example, is one of the most widely used techniques which extracts linear relations that best explain the correlated structure in the data. Following this line of thinking, Zang et al. (2016) proposed a PCA-based approach for bias correction and data integration in high-dimensional genomic data. Meng et al. (2016) also explored the use of dimension reduction techniques to integrate multisource genomic data. The second approach resorts to nonnegative matrix factorization to find meaningful and interpretable features in the data. This is the approach taken by Zhang et al. (2011) to jointly analyze the microRNA (miRNA) and gene expression profiles. Zhang et al. (2012) also employed a joint matrix factorization method to integrate multidimensional

genomic data and identified subsets of messenger RNA (mRNAs), miRNAs, and methylation markers that exhibit similar patterns across different samples or types of measurements. The third approach relies on model-based data integration. For example, Letunic et al. (2004) developed a database for genomic data integration using the hidden Markov model which provided extensive annotations for the domains in the genomic sequence. Ernst and Kellis (2012) proposed an automated learning system for characterizing chromatin states which used the multivariate hidden Markov model to learn chromatin-state assignment for each genomic position from multiple sources of epigenetic data.

Due to the special features of multisource data, there are several challenges in integrating them. First, the multisource data are often collected using different experimental techniques and thus come in a wide variety of formats or data types. In the above case of characterizing epigenomic landscape, histone modification levels are collected using the ChIP sequencing (ChIP-seq) technique and recorded as continuous measurements, whereas DNA methylation levels are collected using a bisulfite sequencing technique and range from 0 to 1. Second, the multisource data are not necessarily the direct measurements of the same physical quantity, even though they are all manifestations of a certain latent variable. For example, ChIP-seq characterizes the epigenetic activity of the protein tails of histone at different amino acid positions, while whole genome bisulfite sequencing (WGBS) provides single-cytosine methylation levels cross the whole genome. Third, the relationship between the multisource measurements and the latent variable is often unknown. For these reasons it is difficult to apply conventional methods, such as the PCA, to perform the integrative analysis.

To address the above issues, we propose a unified framework for integrating data from multiple sources. Assume there exists a latent variable Y, to which the multisource measurements, say $W_1, \ldots, W_K$, are related in some linear or nonlinear ways $f_1^0, \ldots, f_K^0$. That is,

$$f_1^0(W_1) = Y, \ldots, f_K^0(W_K) = Y. \tag{1}$$

We seek to find a representation of the latent variable that captures the common information contained in $f_1^0(W_1), \ldots, f_K^0(W_K)$, even when $f_1^0, \ldots, f_K^0$ are unknown. In epigenetic data, $W_1, \ldots, W_K$ represent the histone modifications and DNA methylation levels collected from multiple sources (Figure 1). We aim to find a representation, the epigenetic index, of the latent variable Y. Intuitively, when $f_1^0, \ldots, f_K^0$ are known functions, Y can be represented by averaging $f_1^0(W_1), \ldots, f_K^0(W_K)$, where each one itself is an estimate of Y. In practice, however, we can hardly know $f_1^0, \ldots, f_K^0$, let alone to estimate $\sum_{k=1}^K f_k^0(W_k)/K$. Instead, we propose to find a set of transformation functions $h_1, \ldots, h_K$ such that the distance between each transformation $h_k(W_k)$ $(k = 1, \ldots, K)$ and their average $\sum_{k=1}^K h_k(W_k)/K$ is minimized. Then, the solution to the above optimization problem provides a group of optimal transformations. Their average naturally can be used as a representation for the latent variable. In other words, we intend to find a one-dimensional representation of the higher dimensional data which is pertinent to a multidimensional scaling problem.

To solve the optimization problem, a technical aspect of our method is to use the basis expansion. It enables the characterization of nonlinear transformations $h_1, \ldots, h_K$ and further simplifies the optimization procedure. With basis expansion the optimal solution takes a specific form which makes our method convenient and efficient to implement. To reflect its relationship with multidimensional scaling and basis expansion, we name our method as B-scaling and the one-dimensional representation as B-mean. The asymptotic property of B-mean is also established to provide theoretical underpinnings for our proposed method. To illustrate its empirical performance, we applied the B-scaling method to integrate multisource epigenetic data and established an epigenetic index (EI) across the genome using

B-mean. The EI showed better explanations of the changes in gene expressions than any of the data source.

The rest of the paper is organized as follows. In Section 2 we introduce the proposed method for integrating multiple measurements. In Section 3 we develop the theoretical properties of the B-mean. Simulation studies and real data analysis are reported in Section 4 and Section 5. Section 6 concludes the paper with a discussion. All the proofs are provided in Appendix.

2. Model setup. In this section, we first consider the population version of the B-scaling method and then generalize to its sample version. Assume that Y is a one-dimensional random variable, which we do not observe. Instead, we observe K numbers, denoted by $W = (W_1, \ldots, W_K)^T \in \mathbb{R}^K$, which measure Y in some linear or nonlinear ways. Our goal is to extract Y from W. In the proposed framework, we assume that the latent variable Y and $W_1, \ldots, W_K$ are connected through some functions $f_1^0, \ldots, f_K^0$. Specifically, let $W_1, \ldots, W_K$ be K random variables and $\mathcal{H}_k$ be a Hilbert space of functions of W_k for $k = 1, \ldots, K$. We assume there exist $f_k^0 \in \mathcal{H}_k$ such that $Y = f_k^0(W_k)$ for $k = 1, \ldots, K$.

To recover the transformations $f_1^0, \ldots, f_K^0$, we can use the fact that, if such transformations do exist, then it must be true that

$$f_1^0(W_1) = \cdots = f_K^0(W_K) = Y.$$

Consequently, $f_1^0, \ldots, f_K^0$ must satisfy the relation

$$\mathrm{E}\left[\sum_{k=1}^{K}\left(f_k^0(W_k) - K^{-1}\sum_{k=1}^{K} f_k^0(W_k)\right)^2\right] = 0.$$

In other words, if $\mathcal{H}_1, \ldots, \mathcal{H}_K$ are families of functions rich enough to contain $f_1^0, \ldots, f_K^0$, then $(f_1^0, \ldots, f_K^0)$ should minimize

$$L(h_1, \ldots, h_K) = \mathrm{E}\left[\sum_{k=1}^{K}\left(h_k(W_k) - K^{-1}\sum_{k=1}^{K} h_k(W_k)\right)^2\right]$$

among $(h_1, \ldots, h_K)$.

However, note that the function is minimized trivially by $(h_1, \ldots, h_K) = (0, \ldots, 0)$ which yields $L(0, \ldots, 0) = 0$. To avoid this trivial solution, we assume that $h_1 \neq 0, \ldots, h_K \neq 0$. Furthermore, since the latent variable Y is unaffected by a multiplicative constant, we assume, without loss of generality, $\|h_1\| = \cdots = \|h_K\| = 1$. Based on these observations, we propose the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad L(h_1, \ldots, h_K), \\ &\text{subject to:} \quad h_1 \in \mathcal{H}_1, \ldots, h_K \in \mathcal{H}_K, \qquad \|h_1\| = \cdots = \|h_K\| = 1. \end{aligned} \tag{2}$$

REMARK 2.1. The ideas of B-scaling and the sigmoid neuron share some common characteristics (Figure 2). A sigmoid neuron takes real-valued data as inputs, and the output function, known as the sigmoid function, is usually a nonlinear function. B-scaling also takes multisource data as inputs. The output functions $h_1, \ldots, h_K$ are defined to be either linear or nonlinear. Furthermore, both methods aim to learn the output functions through minimizing their loss functions. However, B-scaling is performed in an unsupervised manner and allows the output functions to have more general forms. To find those optimal transformations (output functions), B-scaling utilizes the fact that they produce the same output values, which we call as the fused measurement, while the optimization for a sigmoid neuron is implemented in a supervised way.

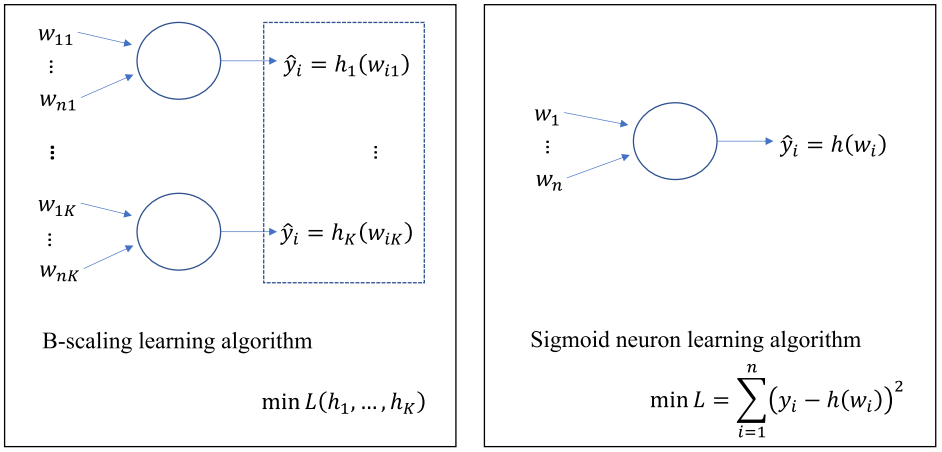

FIG. 2. *B-scaling from the perspective of sigmoid neuron—the building block of deep neural network. For the left panel, w_{ik} is the ith observation for kth measurement, and $\hat{y}_i$ is the ith fused value for $i = 1, \ldots, n$ and $k = 1, \ldots, K$.*

REMARK 2.2 (B-scaling, measurement error model, and multidimensional scaling). Despite its appearance, this is not a measurement error problem (Carroll et al. (2006), Fuller (2009)), because these measurements are not necessarily measuring the same physical quantity. Instead, it is akin to a multidimensional scaling problem (Cox and Cox (2000), Kruskal (1964), Torgerson (1952)): that is, we would like to find a one-dimensional representation of higher dimensional data. However, different from multidimensional scaling, we are not interested in preserving intervariable distances but rather making a connection among several measurements, based on the fact that they are related to the same latent quantity. A technical aspect of our method is the use of basis expansion. Thus, we call our method B-scaling, to reflect its relationship with both multidimensional scaling and basis expansion.

Let $(f_1, \ldots, f_K)$ be the solution to the optimization problem (2). In the ideal situation, where $f_k^0 \in \mathcal{H}_k$ for $k = 1, \ldots, K$, the objective function can be perfectly minimized to 0, and this gives the unique optimizer $(f_1, \ldots, f_K) = (f_1^0, \ldots, f_K^0)$. In this case, any $f_k(W_k)$ can be used as the representation of the latent variable Y. However, in practice we do not make such a strong assumption and allow f_k^0 to be outside of $\mathcal{H}_k$. In this case we use the average of the optimal solution $(f_1, \ldots, f_K)$ as a representation of the latent variable Y. This leads to the notions of B-mean and B-variance, as defined below.

DEFINITION 2.3. Let $f_1, \ldots, f_K$ be the solution to the minimization problem (2). Let

$$\mu_B(W_1, \ldots, W_K) = K^{-1} \sum_{k=1}^{K} f_k(W_k),$$

$$V_B(W_1, \ldots, W_K) = K^{-1} \sum_{k=1}^{K} \big[f_k(W_k) - \mu_B(W_1, \ldots, W_K) \big]^2.$$

We called them B-scaling mean and B-scaling variance, of $W_1, \ldots, W_K$, respectively, or B-mean and B-variance for short.

REMARK 2.4 (B-mean as a representation). Note that the B-mean is a random variable, and its definition does not depend on the assumption that the true transformations $f_k^0 \in \mathcal{H}_k$ for all $k = 1, \ldots, K$. Thus, we have the following two scenarios: *Scenario 1.* When $f_1^0 \in$

$\mathcal{H}_1, \ldots, f_K^0 \in \mathcal{H}_K$, $V_B(W_1, \ldots, W_K)$ is zero, and the unique optimizer $f_1^0(W_1), \ldots, f_K^0(W_K)$ are equal to the B-mean almost surely. The B-mean thus, naturally, becomes a representation of Y up to a multiplicative constant. *Scenario 2.* In a more general case, when some $f_k^0 \notin \mathcal{H}_k$, the distance between B-mean and the latent variable Y is largely determined by the extent to which f_k^0 deviates from $\mathcal{H}_k$. Since we do not make any assumption about the latent variable Y and allow f_k^0 to be outside $\mathcal{H}_k$, the B-mean, as defined here, is only a representation of the latent variable Y to the best of our knowledge of observed measurements $W_1, \ldots, W_K$.

We call B-mean the fused measurement and $W_1, \ldots, W_K$ surrogate measurements. We have to emphasize that we do not intend to estimate the latent measurement Y. Instead, we are interested in extracting common information contained in mutisource data. Similar to the concept of principal component, B-mean is another type of representation of the data such that its distance with all K transformations is minimized. In the rest of the paper, we focus on discussing the properties of B-mean.

For the choice of $\mathcal{H}_1, \ldots, \mathcal{H}_K$, we propose the clans of spline functions, as stated by the condition below. It is convenient to express functions in $\mathcal{S}(m, t^{(k)})$ in terms of spline basis.

CONDITION 2.5. *For each $k = 1, \ldots, K$, let $t^{(k)}$ be the set $\{(t_0^{(k)}, \ldots, t_{k_0}^{(k)}) : 0 = t_0^{(k)} < \cdots < t_{k_0}^{(k)} = 1\}$. Let $\mathcal{S}(m, t^{(k)})$ be the set of spline functions of order m ($m \geq 1$) with knots $t^{(k)}$. That is, for $m = 1$, $\mathcal{S}(m, t^{(k)})$ is the set of step functions with jumps at the knots; for $m = 2, 3, \ldots$,*

$$\mathcal{S}\big(m, t^{(k)}\big) = \big\{s \in C^{m-2}[0, 1] : s(x) \text{ is a polynomial of degree } (m-1) \text{ on each subinterval } \big[t_i^{(k)}, t_{i+1}^{(k)}\big]\big\},$$

where $C^q[0, 1]$ is the clan of functions on $[0, 1]$ with continuous qth derivative. We assume $\mathcal{H}_k = \mathcal{S}(m, t^{(k)})$.

For any fixed m and $t^{(k)}$, let $\{N_{1,m}^{(k)}(x), \ldots, N_{\ell,m}^{(k)}(x)\}$ be the basis of $\mathcal{S}(m, t^{(k)})$, where $\ell = k_0 + m - 1$ is the number of basis functions in $\mathcal{S}(m, t^{(k)})$. For any $h_k(x) \in \mathcal{S}(m, t^{(k)})$, there exits $A_k \in \mathbb{R}^\ell$ such that $h_k(x) = A_k^T \mathbf{N}_k(x)$, where $\mathbf{N}_k(x) = (N_{1,m}^{(k)}(x), \ldots, N_{\ell,m}^{(k)}(x))^T \in \mathbb{R}^\ell$. By Condition 2.5, for $k = 1, \ldots, K$,

$$h_k(W_k) = A_k^T \mathbf{N}_k(W_k).$$

The objective function in (2) can then be written in matrix as

$$\mathrm{E}\left[\sum_{k=1}^K \left(A_k^T \mathbf{N}_k(W_k) - K^{-1} \sum_{k=1}^K A_k^T \mathbf{N}_k(W_k)\right)^2\right], \quad A_1, \ldots, A_K \in \mathbb{R}^\ell. \tag{3}$$

Let $A = (A_1^T, \ldots, A_K^T)^T$, $\mathbb{N} = \operatorname{diag}(\mathbf{N}_1(W_1), \ldots, \mathbf{N}_K(W_K))$, $\mathbf{1} = (1, 1, \ldots, 1)^T \in \mathbb{R}^K$, and $Q = I - \mathbf{1}\mathbf{1}^T/K$. Then, the objective function $L(A_1, \ldots, A_K)$ can be rewritten as

$$L(A) = \mathrm{E}\|\mathbb{N}^T A - K^{-1}\mathbf{1}\mathbf{1}^T \mathbb{N}^T A\|^2 = A^T \mathrm{E}\big(\mathbb{N} Q \mathbb{N}^T\big) A.$$

Thus, at the population level, optimization (2) amounts to solving the following generalized eigenvalue problem:

$$\begin{aligned} &\text{minimize} \quad A^T \mathrm{E}\big(\mathbb{N} Q \mathbb{N}^T\big) A, \\ &\text{subject to:} \quad A^T \operatorname{var}(\mathbb{N}\mathbf{1}/K) A = 1. \end{aligned} \tag{4}$$

Let $\Lambda = \mathrm{E}(\mathbb{N} Q \mathbb{N}^T)$ and $\Sigma = \operatorname{var}(\mathbb{N}\mathbf{1}/K)$, the above generalized eigenvalue problem yields the following explicit expression of the B-mean.

Algorithm 1 B-scaling Algorithm

1. Rescale $\{w_{i1}, \ldots, w_{iK}\}_{i=1}^n$ such that they are all defined on $[0, 1]$. For kth measurement $\{w_{ik}\}_{i=1}^n$ $(k = 1, \ldots, K)$, generate ℓ spline functions of order m. Denote the spline functions as $N_{1,m}^{(k)}(x), \ldots, N_{\ell,m}^{(k)}(x)$.
2. Evaluate the spline functions at each data point of the K measurements, resulting in

$$\mathbf{N}_1(w_{i1}) = \begin{pmatrix} N_{1,m}^{(1)}(w_{i1}) \\ \vdots \\ N_{\ell,m}^{(1)}(w_{i1}) \end{pmatrix}, \ldots, \mathbf{N}_K(w_{iK}) = \begin{pmatrix} N_{1,m}^{(K)}(w_{iK}) \\ \vdots \\ N_{\ell,m}^{(K)}(w_{iK}) \end{pmatrix}, \quad i = 1, \ldots, n.$$

Let $\mathbb{N}_i = \operatorname{diag}(\mathbf{N}_1(w_{i1}), \ldots, \mathbf{N}_K(w_{iK}))$. Calculate Λ_n and Σ_n by equation (5).
3. Find the eigenvector of $\Sigma_n^{-1/2}\Lambda_n\Sigma_n^{-1/2}$ corresponding to its smallest eigenvalue, denoted as $\hat{\mathbf{b}}$. Output $\hat{\boldsymbol{a}} = \Sigma_n^{-1/2}\hat{\mathbf{b}}$.
4. Estimate the B-mean at the data points $(w_{i1}, \ldots, w_{iK})$ by $\hat{\mu}_B(w_{i1}, \ldots, w_{iK}) = \hat{\boldsymbol{a}}^T\mathbb{N}_i\mathbf{1}/K$ for $i = 1, \ldots, n$.

THEOREM 2.6. *The B-mean of $W_1, \ldots, W_K$ has the form of*

$$\mu_B(W_1, \ldots, W_K) = \boldsymbol{a}^T\mathbb{N}\mathbf{1}/K,$$

where $\boldsymbol{a} = \Sigma^{-1/2}\mathbf{b}$, and $\mathbf{b}$ is the eigenvector of $\Sigma^{-1/2}\Lambda\Sigma^{-1/2}$ corresponding to its smallest eigenvalue.

At the sample level, assume that we observe i.i.d. samples $\mathbf{w}_i = (w_{i1}, \ldots, w_{iK})^T, i = 1, \ldots, n$ which measures the i.i.d. latent variable y_i in some linear or nonlinear ways. We use the sample-level counterpart of $\mu_B(W_1, \ldots, W_K)$ in Theorem 2.6 to estimate the B-mean. Specifically, let

$$\begin{aligned} &\mathbb{N}_i = \operatorname{diag}\big(\mathbf{N}_1(w_{i1}), \ldots, \mathbf{N}_K(w_{iK})\big), \\ &\Lambda_n = n^{-1}\sum_{i=1}^n \mathbb{N}_i Q\mathbb{N}_i^T, \\ &\Sigma_n = n^{-1}\sum_{i=1}^n \left(\mathbb{N}_i\mathbf{1} - n^{-1}\sum_{i=1}^n \mathbb{N}_i\mathbf{1}\right)\left(\mathbb{N}_i\mathbf{1} - n^{-1}\sum_{i=1}^n \mathbb{N}_i\mathbf{1}\right)^T/K^2. \end{aligned} \tag{5}$$

The estimate of $\mu_B(w_{i1}, \ldots, w_{iK})$ is

$$\hat{\mu}_B(w_{i1}, \ldots, w_{iK}) = \hat{\boldsymbol{a}}^T\mathbb{N}_i\mathbf{1}/K,$$

where $\hat{\boldsymbol{a}} = \Sigma_n^{-1/2}\hat{\mathbf{b}}$, and $\hat{\mathbf{b}}$ is the eigenvector of $\Sigma_n^{-1/2}\Lambda_n\Sigma_n^{-1/2}$ corresponding to its smallest eigenvalue. In Algorithm 1 we show an algorithm for computing the B-mean based on the data $(w_{i1}, \ldots, w_{iK}), i = 1, \ldots, n$.

3. Theoretical properties. In this section we study the asymptotic property of the B-mean. Recall that the B-mean at $(w_{i1}, \ldots, w_{iK})$ is estimated by $\hat{\mu}_B(w_{i1}, \ldots, w_{iK}) = \hat{\boldsymbol{a}}^T\mathbb{N}_i\mathbf{1}/K$, where $\hat{\boldsymbol{a}} = \Sigma_n^{-1/2}\hat{\mathbf{b}}$ and $\hat{\mathbf{b}}$ is the eigenvector of $\Sigma_n^{-1/2}\Lambda_n\Sigma_n^{-1/2}$ corresponding to its smallest eigenvalue. To show the asymptotic distribution of the B-mean, we first show that both $\hat{\mathbf{b}}$ and $\hat{\boldsymbol{a}}$ are asymptotically normally distributed. To simplify presentation, we introduce the following notations. Let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the maximum and minimum eigenvalues/singular values of a matrix, respectively. Let $M_1 \otimes M_2$ denote the Kronecker product

of two matrices M_1 and M_2. Let M^+ denote the pseudo-inverse of a matrix M. Let $\xrightarrow{D}$ denote in distribution convergence and $\xrightarrow{P}$ in probability convergence. We make the following assumptions.

CONDITION 3.1. *Without loss of generality, we assume that* $\{w_{i1}, \ldots, w_{iK}\}_{i=1}^n$ *are i.i.d. random variables in* $[0, 1]$. *Moreover, we assume that each* w_{ik} *has a finite* $4(m-1)$*th moment.*

CONDITION 3.2. $0 < \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) < \infty$.

The finite $4(m-1)$th moment assumption in Condition 3.1 guarantees that, when the basis $\{\mathbf{N}_1(x), \ldots, \mathbf{N}_K(x)\}$ are of mth order, the random sequence $n^{-1}\sum_{i=1}^n \times N_{\ell_1,m}^{k_1}(w_{ik_1})N_{\ell_2,m}^{k_2}(w_{ik_2})$ has an asymptotic normal distribution, and it is true for all $k_1, k_2 = 1 \ldots, K$ and $\ell_1, \ell_2 = 1, \ldots, \ell$. The asymptotic properties of Λ_n and Σ_n thus follow. Condition 3.2 ensures that no $N_{\ell_1,m}^{(k_1)}(w_{ik_1})$ has a dominate variance or is linearly dependent on $N_{\ell_2,m}^{(k_2)}(w_{ik_2})$ for $k_1 \neq k_2$ and $\ell_1 \neq \ell_2$. It also implies the existence of Σ^{-1}. In the following we denote $\Sigma^{-1/2}\Lambda\Sigma^{-1/2}$ by R, $\Sigma_n^{-1/2}\Lambda_n\Sigma_n^{-1/2}$ by R_n, and $\mathbb{N}\mathbf{1}/K$ by $\mathbb{Z}$. With Conditions 3.1 and 3.2, we have the following asymptotic distribution for R_n.

THEOREM 3.3. *Under Conditions 3.1–3.2, we have*

$$\sqrt{n}\big[\operatorname{vec}(R_n - R)\big] \xrightarrow{D} N(0, \Pi_R), \tag{6}$$

where

$$\Pi_R = \begin{pmatrix}\Omega_1 & \Omega_2\end{pmatrix}\begin{pmatrix}\Phi_{11} & \Phi_{12}\\ \Phi_{21} & \Phi_{22}\end{pmatrix}\begin{pmatrix}\Omega_1^T\\ \Omega_2^T\end{pmatrix}$$

and

$$\begin{aligned}
\Omega_1 &= -\big(\Sigma^{-\frac{1}{2}}\Lambda \otimes I + I \otimes \Sigma^{-\frac{1}{2}}\Lambda\big)\big(\Sigma \otimes \Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}} \otimes \Sigma\big)^{-1},\\
\Omega_2 &= \big(\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}}\big),\\
\Phi_{11} &= \mathrm{E}\big\{\big[\mathbb{Z}\otimes\mathbb{Z} - \mathrm{E}(\mathbb{Z}\otimes\mathbb{Z}) - \big(\mathbb{Z} - \mathrm{E}(\mathbb{Z})\big)\otimes\mathrm{E}(\mathbb{Z}) - \mathrm{E}(\mathbb{Z})\otimes\big(\mathbb{Z} - \mathrm{E}(\mathbb{Z})\big)\big]\\
&\qquad \times \big[\mathbb{Z}\otimes\mathbb{Z} - \mathrm{E}(\mathbb{Z}\otimes\mathbb{Z}) - \big(\mathbb{Z} - \mathrm{E}(\mathbb{Z})\big)\otimes\mathrm{E}(\mathbb{Z}) - \mathrm{E}(\mathbb{Z})\otimes\big(\mathbb{Z} - \mathrm{E}(\mathbb{Z})\big)\big]^T\big\},\\
\Phi_{21} &= \mathrm{E}\big\{\big[(\mathbb{N}\otimes\mathbb{N})\operatorname{vec}(Q) - \mathrm{E}(\mathbb{N}\otimes\mathbb{N})\operatorname{vec}(Q)\big]\\
&\qquad \times \big[\mathbb{Z}\otimes\mathbb{Z} - \mathrm{E}(\mathbb{Z}\otimes\mathbb{Z}) - \big(\mathbb{Z} - \mathrm{E}(\mathbb{Z})\big)\otimes\mathrm{E}(\mathbb{Z}) - \mathrm{E}(\mathbb{Z})\otimes\big(\mathbb{Z} - \mathrm{E}(\mathbb{Z})\big)\big]^T\big\},\\
\Phi_{22} &= \mathrm{E}\big\{\big[(\mathbb{N}\otimes\mathbb{N})\operatorname{vec}(Q) - \mathrm{E}(\mathbb{N}\otimes\mathbb{N})\operatorname{vec}(Q)\big]\big[(\mathbb{N}\otimes\mathbb{N})\operatorname{vec}(Q) - \mathrm{E}(\mathbb{N}\otimes\mathbb{N})\operatorname{vec}(Q)\big]^T\big\}.
\end{aligned}$$

In the rest of the paper, we write the matrix

$$\begin{pmatrix}\Phi_{11} & \Phi_{12}\\ \Phi_{21} & \Phi_{22}\end{pmatrix}$$

in Theorem 3.3 as Φ. To further investigate the properties of the eigenvectors of R_n, we need the following condition.

CONDITION 3.4. *Assume that matrix* $R \in \mathbb{R}^{K\ell\times K\ell}$ *has distinct eigenvalues, denoted as* $d_1, \ldots, d_r$, *such that* $d_1 > \cdots > d_r > 0$, *where* $r = K\ell$.

Condition 3.4 requires that the eigenvalues of R are separable which is to guarantee that the eigenvectors are uniquely defined. To conclude, we transform the minimization problem (2) to the generalized eigenvalue problem (4) with Condition 2.5, and the optimal solution to (4) is the eigenvector corresponding to the smallest eigenvalue of the matrix $R = \Sigma^{-1/2}\Lambda\Sigma^{-1/2}$. When Conditions 3.2 and 3.2 are further satisfied, the matrix R is invertible and has separable eigenvalues which ensures the uniquely defined eigenvectors. Thus, there exists a unique solution to the minimization problem (2), given Conditions 2.5, 3.2, and 3.4. The next lemma gives the asymptotic distributions of $\hat{\boldsymbol{a}}$ and $\hat{\mathbf{b}}$.

LEMMA 3.5. *Under Conditions 3.1–3.4, we have*

$$\sqrt{n}(\hat{\mathbf{b}} - \mathbf{b}) \xrightarrow{D} N\big(\mathbf{0}, (M_{b1}, M_{b2})\Phi(M_{b1}, M_{b2})^T\big),$$

where Φ is defined in Theorem 3.3,

$$M_{b1} = -\big[\mathbf{b}^T \otimes V(d_r I - \Gamma)^+ V^T\big]\big(\Sigma^{-\frac{1}{2}}\Lambda \otimes I + I \otimes \Sigma^{-\frac{1}{2}}\Lambda\big)\big(\Sigma \otimes \Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}} \otimes \Sigma\big)^{-1},$$
$$M_{b2} = \big[\mathbf{b}^T \otimes V(d_r I - \Gamma)^+ V^T\big]\big(\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}}\big)$$

with Γ being the diagonal matrix $\operatorname{diag}(d_1, \ldots, d_r)$*, and V being a matrix whose columns are eigenvectors of R. The asymptotic distribution of $\hat{\boldsymbol{a}}$ is given by*

$$\sqrt{n}(\hat{\boldsymbol{a}} - \boldsymbol{a}) \xrightarrow{D} N\big(\mathbf{0}, (M_{a1}, M_{a2})\Phi(M_{a1}, M_{a2})^T\big),$$

where

$$M_{a1} = -\big(\mathbf{b}^T \otimes I\big)\big(\Sigma \otimes \Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}} \otimes \Sigma\big)^{-1}$$
$$- \Sigma^{-\frac{1}{2}}\big[\mathbf{b}^T \otimes V(d_r I - \Gamma)^+ V^T\big]\big(\Sigma^{-\frac{1}{2}}\Lambda \otimes I + I \otimes \Sigma^{-\frac{1}{2}}\Lambda\big)\big(\Sigma \otimes \Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}} \otimes \Sigma\big)^{-1},$$
$$M_{a2} = \Sigma^{-\frac{1}{2}}\big[\mathbf{b}^T \otimes V(d_r I - \Gamma)^+ V^T\big]\big(\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}}\big).$$

As showed in Lemma 3.5, $\sqrt{n}(\hat{\boldsymbol{a}} - \boldsymbol{a})$ is asymptotically normally distributed. Notice that for a given new observation, denoted as $(\tilde{w}_1, \ldots, \tilde{w}_K)$, its B-mean estimate is represented by an average of $\mathbf{N}_1^T(\tilde{w}_1)\hat{\boldsymbol{a}}_1, \ldots, \mathbf{N}_K^T(\tilde{w}_K)\hat{\boldsymbol{a}}_K$, where $\mathbf{N}_1^T(\tilde{w}_1), \ldots, \mathbf{N}_K^T(\tilde{w}_K)$ are the values of basis functions evaluated at $(\tilde{w}_1, \ldots, \tilde{w}_K)$. With Lemma 3.5 we show in Theorem 3.6 that the B-mean estimate follows asymptotic normal distribution.

THEOREM 3.6. *Suppose that Conditions 3.1 through 3.4 are satisfied. For a given new observation $\tilde{w} = (\tilde{w}_1, \ldots, \tilde{w}_K)^T$, let $\mu_B(\tilde{w})$ be the B-mean at $\tilde{w}$, and let $\hat{\mu}_B(\tilde{w})$ be the B-mean estimate. Then, as $n \to \infty$,*

$$\sqrt{n}\big[\hat{\mu}_B(\tilde{w}) - \mu_B(\tilde{w})\big] \xrightarrow{D} N\big(\mathbf{0}, \sigma_\mu^2\big), \tag{7}$$

where

$$\sigma_\mu^2 = \mathbf{N}(\tilde{w})^T(M_{a1}, M_{a2})\Phi(M_{a1}, M_{a2})^T\mathbf{N}(\tilde{w})/K^2$$

with $\mathbf{N}(\tilde{w})$ being $(\mathbf{N}_1^T(\tilde{w}_1), \ldots, \mathbf{N}_K^T(\tilde{w}_K))^T$.

4. Simulation studies. We have conducted comprehensive simulation studies to investigate the empirical performance of the proposed B-mean for representing the underlying latent variable. In the first example we discussed some implementation issues of the B-scaling method. In the second example we examined the performance of the B-mean by comparing with that of principal components (PCs) and multidimensional scaling (MDS). Their performances were evaluated by their correlations with the latent measurement. For principal component analysis we reported the maximum correlations that all the PCs can achieve ($\mathrm{PC}_{\max}$). For multidimensional scaling we choose to map all the data points to a one-dimensional space.

EXAMPLE 1. The selection of basis function, its order, and the knots are key issues in implementing the B-scaling algorithm. In practice, we choose the natural spline of order $m = 4$. The flexibility of a spline then mainly lies in the selection of knots. We let knots be the quantiles of $(w_{1k}, \ldots, w_{nk})$ for $k = 1, \ldots, K$. In this paper we focus on the B-scaling method itself and choose the same values of m and k_0 for all K different measurements. Then, for different $k = 1, \ldots, K$, the flexibility of fitting splines only lies in the difference of their quantiles. For further studies we can embed into our algorithm the data-driven methods for fitting splines (He, Shen and Shen (2001), Yuan, Chen and Zhou (2013)).

In this example we demonstrate how the number of knots impacts the behavior of the B-scaling method. To generate multiple nonlinear functions automatically, we consider the Logit function

$$g(x) = \frac{1}{1 + \exp[20(x - 0.5)]}$$

and let $(w_{i1}, \ldots, w_{iK})$ measure y_i in nonlinear ways, $w_{ik} = \sum_{t=1}^{H} s_k Z_{kt} \delta_t g(y_i + \epsilon_{ik})$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. We generated y_i from $U(0, 1)$ as the latent measurement and the errors ϵ_{ik} from $N(0, 0.1)$. In addition, s_k is a scale parameter generated from $U(-10, 10)$, Z_{kt} from uniform distribution $U(-\sqrt{3}, \sqrt{3})$, and $\delta_t = (-1)^{t+1} t^{-\nu/2}$. Parameters ν and H were set to 2 and 5, respectively. Let $n = 1000, 2000, 3000$, $K = 10, 20, 30$, and k_0 vary from 11 to 25. The performances of B-scaling mean were evaluated by its correlation with y_i. As illustrated in Figure 3, under different scenarios the estimated B-means maintained high correlations with y_i, as we vary the number of knots. In practice, we recommend to utilize the B-variance to select a proper number of knots. That is, we propose to choose the number of knots that generates a smaller B-variance. Intuitively, a proper number of knots would result in a better estimation of the K transformations $f_1, \ldots, f_K$ and further a smaller B-variance among $f_1(W_1), \ldots, f_K(W_K)$. We have implemented this method for selecting the number of knots in simulation studies. In the Supplementary Material (Liu et al. (2022)) we further discussed the possibility of combining the plot of B-variance and the associated eigensystem to determine the number of knots (Supplementary Material, Figure S1) for future studies.We also reported the computation time of B-scaling in Figure 4. It only takes seconds to calculate the B-mean when the sample size and the number of measurements are relatively small (Figure 4, left panel). As the sample size n, number of measurements K, and the number of knots increase, the computation time of B-scaling tends to increase (Figure 4).

EXAMPLE 2. In this example we investigated the performances of the B-mean for representing the latent variable. We generated y_i as the latent measurement under two scenarios: (1) uniform distribution $U(0, 1)$ and (2) normal distribution $N(0, 1)$. The K observed measurements $(w_{i1}, \ldots, w_{iK})$ as for $i = 1, \ldots, n$ were simulated using the same model in Example 1. Let $n = (500, 700, 1000, 2000, 3000)$ and $K = (7, 10, 20, 30)$. We generated 100

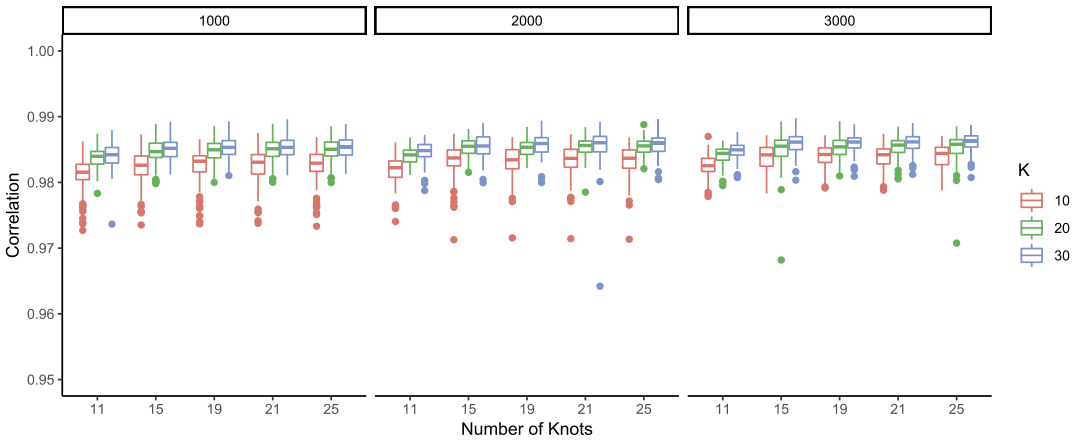

FIG. 3. *Boxplots of the correlations between the latent variable y_i and estimated B-means.*

datasets for each combination of n and K. To evaluate the performances of the B-mean and PCs, their correlations with y_i in each simulation are reported. The quantities

$$\rho_{\max} = \max_k |\mathrm{Corr}(w_{ik}, y_i)|, \qquad \bar{\rho}_0 = K^{-1} \sum_{k=1}^{K} |\mathrm{Corr}(w_{ik}, y_i)|$$

are also recorded. Figure 5 displays the boxplots of the correlations between the latent variable y_i and the aforementioned estimates based on 100 simulation runs. The performances of $\rho_{\max}$ and $\bar{\rho}_0$ illustrate that w_{ik}s have relatively high correlations with y_i on average. For principal component analysis we reported the maximum correlations that all the PCs can achieve. However, since the PC is a linear combination of the observed measurements $(w_{i1}, \ldots, w_{iK})$, its correlation with y_i thus heavily depends on whether w_{ik} is linearly correlated with y_i. In all scenarios the B-mean has the highest correlation on average with the latent variable y_i.

EXAMPLE 3. To show the performance of the B-scaling method when g functions have different structures, we consider

$$g(x) = \frac{1}{1 + \exp[20(x - 0.5)]} \quad \text{and} \quad g_t(x) = \log(|t/x|)$$

and let $(w_{i1}, \ldots, w_{iK})$ measure y_i in nonlinear ways, where $w_{ik} = \sum_{t=1}^{H} s_k Z_{kt} \delta_t g(y_i + \epsilon_{ik})$ for $k = 1, \ldots, \lceil K/2 \rceil$ and $w_{ik} = \sum_{t=1}^{H} s_k Z_{kt} \delta_t g_t(y_i + \epsilon_{ik})$ for $k = \lceil K/2 \rceil + 1, \ldots, K$. Other

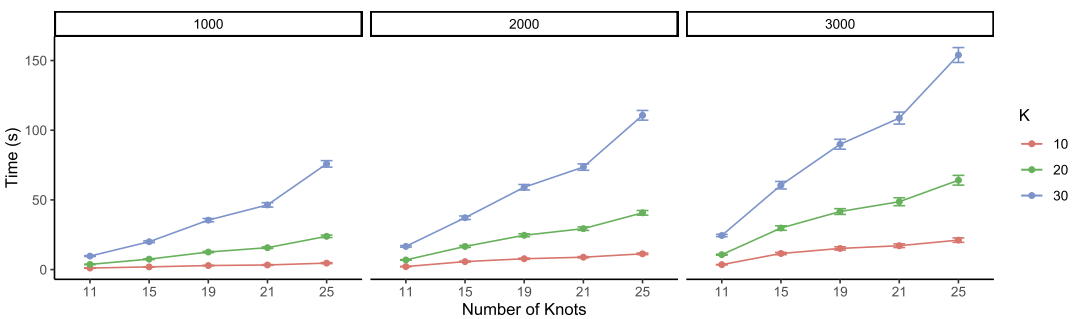

FIG. 4. *The computation time of B-scaling method with varying n, K, and the number of knots. The error bar represents one standard deviation away from the mean.*

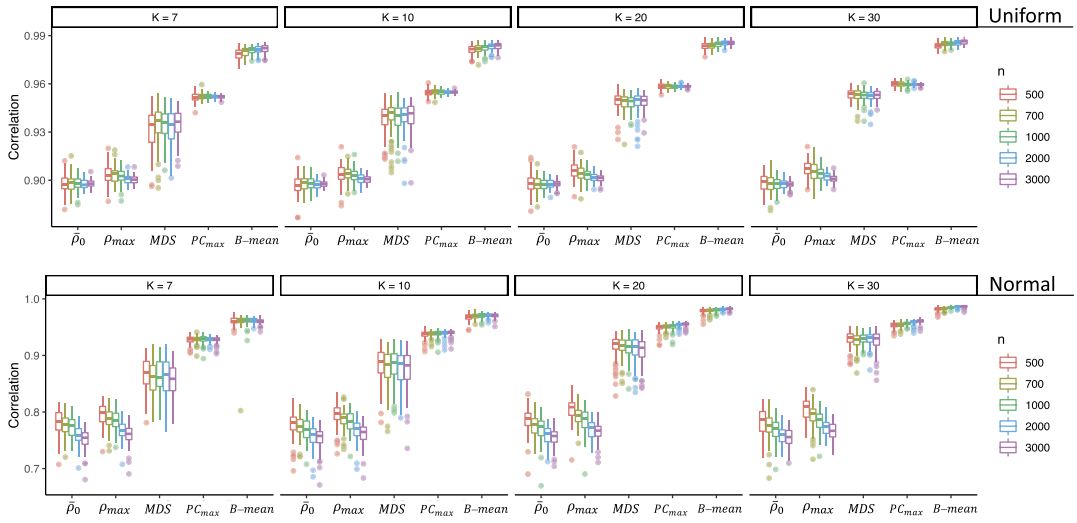

FIG. 5. *Boxplots of the correlations between the latent variable y_i and the aforementioned estimates in Example 2. The latent variable is generated from uniform distribution $U(0, 1)$ (upper panel) and standard normal distribution $N(0, 1)$ (lower panel).*

settings are the same as Example 2. In all scenarios, B-scaling method outperforms its competitors (Figure 6). We further increased the variance of ϵ_{ik} to 0.3 and generated data pairs with the same settings in Examples 2 and 3, respectively. The average correlation $\bar{\rho}_0$ then decreased to a value around 0.65 in the uniform setting (Supplementary Material Figure S2, A and C) and around 0.4 in the normal setting (Supplementary Material Figure S2, B and D). Our method still achieved averaged correlations around 0.9 and 0.7 under each setting, respectively.

5. Application: Epigenetic index (EI). Epigenetic modifications, such as DNA methylation and histone modifications, play an important role in regulating gene expression and numerous cellular processes (Portela and Esteller (2010)). The modifications target on genetic

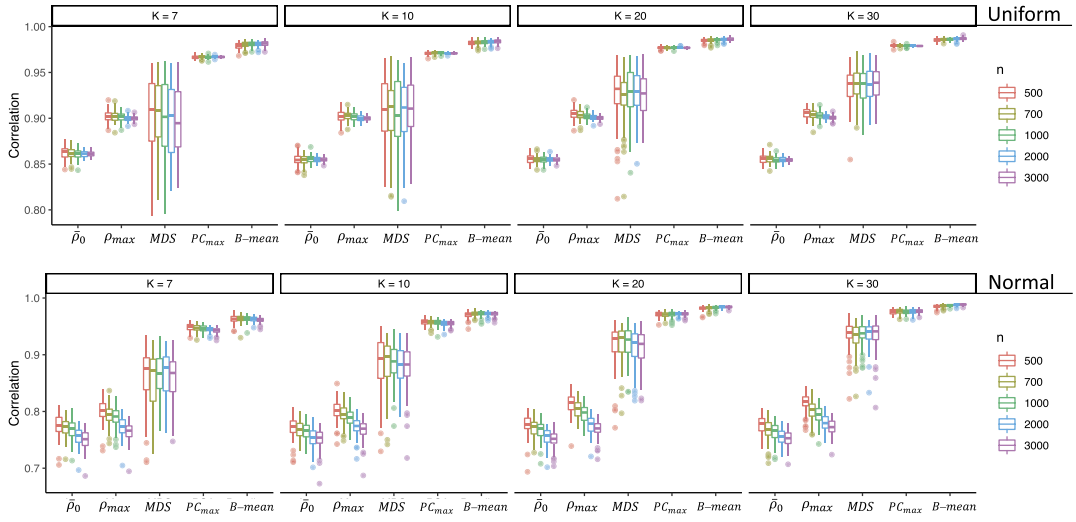

FIG. 6. *Performance comparisons between B-mean and its competitors in Example 3. The latent variable is generated from uniform distribution $U(0, 1)$ (upper panel) and standard normal distribution $N(0, 1)$ (lower panel).*

TABLE 1
Model performance (adjusted R^2) for different methods

Predictor	B-mean	PC_{max}	MDS	Single	All
Adj.R^2	**0.321**	0.214	0.126	0.242	0.304

materials in cells and regulate the biological processes without altering the DNA sequence. Recently, next generation high-throughput sequencing technologies have been adopted to characterize the epigenomic landscapes of primary human tissues and cells (Reuter, Spacek and Snyder (2015)). The resulting datasets, for instance, the Roadmap Epigenomics database, have provided information on various types of histone modifications and DNA methylation levels across the genome. The integration of these different measures of epigenetic information, also known as the construction of chromatin state maps (Ernst and Kellis (2012)), for different tissues and cell types under varied conditions can help researchers understand biological processes, such as morphogenesis, cell differentiation, and tumorigenesis (Feinberg, Koldobskiy and Göndör (2016), Sheaffer et al. (2014)). In this application the epigenetic activeness is latent and cannot be measured directly. We intend to integrate various types of histone modifications and DNA methylation levels to infer epigenetic activeness across the genome.

We used the human liver cancer cell line data (The ENCODE Project Consortium (2012)). The data consists of the information of six types of histone modifications and whole genome DNA methylation levels measured by ChIP-seq and WGBS, respectively. These six histone modifications, that is, methylation and acetylation to histone proteins H3 at different amino acid positions, include H3K9ac, H3K27ac, H3K4me1, H3K4me2, H3K9me3, and H3K27me3. We used the values of fold change over control for histone modifications and DNA methylation levels on the promoter region, that is, 1000 base pairs upstream or downstream of where a gene's transcription begins. After filtering the genes with zero histone modification and DNA methylation levels, we selected 4293 genes. In summary, the kth observed measurement w_{ik} here represents kth epigenetic modification level of promoter region for ith gene, where $k = 1, \ldots, 7$, and $i = 1, \ldots, 4293$. We aim to integrate these seven epigenetic modifications to infer the latent epigenetic activeness y_i. The proposed method, PCA, and MDS were applied to the dataset. To compare the performances of these methods, we considered the following linear regression model and investigated to what extent the fused chromatin EI calculated by these two methods could explain the gene expression levels:

$$\log(g_i) = \alpha_0 + \alpha_1 x_i + \epsilon_i, \tag{8}$$

where g_i is the expression level (reads per kilobase million, RPKM) for gene i, $i = 1, \ldots 4293$, and x_i is the fused EI values corresponding to ith gene. The adjusted R^2s of the model (8) for B-mean, PC_{max}, MDS were 0.321, 0.214, and 0.126, respectively, and the full model with all the predictors, that is, six histone modifications and DNA methylation, had adjusted R^2 of 0.304 (Table 1). In addition, the maximum R^2 was 0.242 if we only included one of these seven predictors in the model. The B-mean achieves the highest R^2, indicating that the EI fused using B-scaling approach can explain better the changes in gene expression. Since we used the whole genome data and gene expression was influenced by many other factors, the R^2 around 0.3 was expected (Yuan et al. (2006)).

To further verify the performance of the proposed method, we performed the gene ontology analysis. Through Livercome database (Lee et al. (2011)), a database for liver cancer genes, we downloaded 3660 cancer-related genes. Among these cancer genes, 768 genes were shown in our selected gene list of size 4293. We then divided our gene list into 10

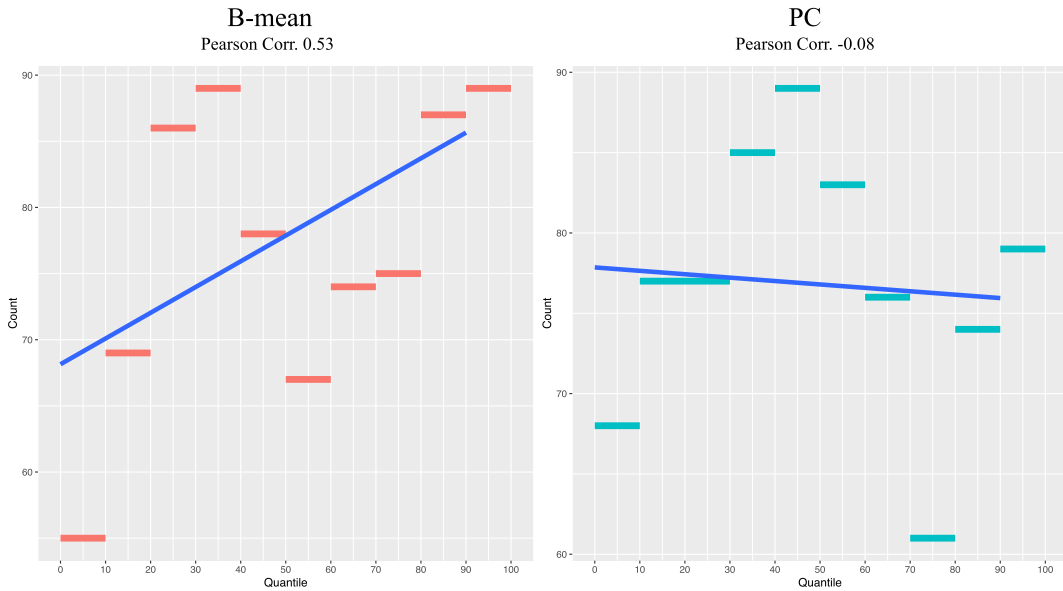

FIG. 7. *The relationship between the number of liver cancer-related genes and quantiles of B-mean (left panel) or* $PC_{\max}$ *(right panel) values. The y-axis is the number of liver cancer-related genes, and the x-axis is the corresponding quantiles. The solid blue lines are fitted using the simple linear regression.*

subsets based on 10 quantiles, which were equally spaced from zero to one hundred, of B-mean or PC values. As $PC_{\max}$ outperforms MDS, we only compared the proposed method with $PC_{\max}$. In Figure 7 we showed how the number of cancer-related genes from Livercome database in each subset changed as the B-mean (left panel) or $PC_{\max}$ (right panel) values increased. The proposed method showed clear advantage over $PC_{\max}$. When the B-mean values grew, the number of cancer-related genes also increased. This indicated that we could find more cancer-related genes when the genes had higher B-mean values, that is, higher EI. The Pearson correlation coefficients in Figure 7 also indicated that the B-mean values and the number of liver cancer genes had a stronger linear trend, compared to the PCA approach.

6. Discussion. The B-scaling approach provides a flexible framework for integrating multisource data. It adopts a basis expansion step to approximate the nonlinear transformations of $W_1, \ldots, W_K$, and an optimization step to search for a one-dimensional surrogate of those measurements. The basis expansion step enables the characterization of nonlinear transformations, while the optimization step, borrowing strength from the PCA approach, is both convenient and efficient to implement. As a trade-off, the B-scaling approach imposes a few assumptions to ensure its efficiency and effectiveness in data fusion. In this approach we only consider a fixed number of basis, order, and number of knots, as stated in Condition 2.5. Condition 2.5 is only a working assumption. It simplifies the optimization process and yields a nice and simple optimal solution, although such an assumption is restrictive. It is possible to relax Condition 2.5 and allow the number of knots to grow with the sample size. The asymptotic properties of B-mean then may depend on the order between the number of knots and the sample size. Another way to make the method fully nonparametric is to impose penalties to the functions $h_1, \ldots, h_K$. The objective function thus becomes

$$L(h_1, \ldots, h_K) = \mathrm{E}\left[\sum_{k=1}^{K}\left(h_k(W_k) - K^{-1}\sum_{k=1}^{K} h_k(W_k)\right)^2\right] + \lambda \sum_{k=1}^{K} \int h_k''(x)^2\, dx,$$

which may further be simplified to have the following quadratic form:

$$L(h_1, \ldots, h_K) = A^T \left[\mathrm{E}(\mathbb{N} Q \mathbb{N}^T) + \lambda D\right] A,$$

where D is a block diagonal matrix with kth diagonal element be D_k, and D_k is a ℓ-by-ℓ kernel matrix that measures the roughness of h_k. For any given smoothing parameter λ, this optimization problem becomes an eigenvalue problem. These ideas deserve to be further explored in future studies. Regarding the number of knots, it should depend on both the sample size and the true transformations between the latent variable and the observed measurements. With larger sample size and more complex transformations, we recommend to use more knots. While the suggestion for the number of knots in our manuscript is only a naive recommendation, it works well under various scenarios. In this paper we propose to choose the number of knots with a smaller B-variance.

The robustness of the B-scaling method is another issue. If one of the K measurements severely deviates from the true measurement, it would have an impact on the minimization process. We could use the following ideas to improve the robustness of the B-scaling method. First, the B-variance can be used to identify a potential poor measurement, because removing it would result in a dramatic decreasing in the B-variance. Thus, if we are able to obtain the information about the accuracy of each measurement, we may put weights on the measurements accordingly to down weight poor measurements. The objective function can then be set as $\mathrm{E}[\sum_{k=1}^{K} \delta_k (h_k(W_k) - K^{-1} \sum_{k=1}^{K} h_k(W_k))^2]$, where δ_k is the weight for measurement k. Second, to make this optimization process more robust, we could also change the L_2-norm to L_1-norm in the objective function which deserves further investigation.

Regarding the type of transformation functions, we choose to treat the monotone transformations as a special case for the B-scaling method rather than explicitly imposing it as a constraint. If the transformation for W_k is indeed monotone, then the optimal procedure would pick up this information automatically. Since the B-scaling method is based on a minimization problem that is transformed into an eigenvalue problem, imposing an extra constraint would disrupt this simplicity and add significant complexity to the computation with only limited benefit and efficiency. For a more refined version of B-scaling, one might impose some inequality constraints on the spline coefficients to control the trend of the transformations.

Currently, we only considered W_k $(k = 1, \ldots, K)$ as a univariate random variable and proposed using the B-scaling mean as a representation of the K different measurements. When it comes to integrating vector-valued samples, we could generalize W_k to be a p-dimensional random vector, meaning that p features for each subject are observed in measurement k. Correspondingly, we may reformulate the objective function as

$$L(h_1, \ldots, h_K) = \mathrm{E}\left[\sum_{k=1}^{K} \left\| h_k(W_k) - K^{-1} \sum_{k=1}^{K} h_k(W_k) \right\|_2^2\right],$$

where $W_k = (W_{k1}, \ldots, W_{kp})^T \in \mathbb{R}^p$ and each h_k is a multivariate smooth function. The multivariate smooth function can be constructed through tensor product smoothing. For instance, if $p = 3$, then h_k can be represented as

$$h_k(W_k) = \sum_{ijl} \beta_{ijl} B_{1i}(W_{k1}) B_{2j}(W_{k2}) B_{3l}(W_{k3}),$$

where the β_{ijl} are coefficients and $B_{1i}(W_{k1})$, $B_{2j}(W_{k2})$, and $B_{3l}(W_{k3})$ are the basis functions. The objective function can be further simplified to have a quadratic form. Such a generalization is an interesting topic for future investigation.

In summary, we believe that the B-scaling method has a broad and important impact on applications in many areas. To facilitate the method development in this direction, we implemented the B-scaling algorithm using programming language R and which is available in Supplement B.

APPENDIX: PROOF

In this Appendix we prove the asymptotic results in Section 3. We rely on what is known as the von Mises expansion to perform this task. For a comprehensive treatment on this topic, see Fernholz (2012). The same approach was used in Li and Wang (2007); see also Li (2018).

Let $\mathcal{D}$ be the clan of all distributions of W. Let F_n be the empirical distribution of W, and let F_0 be the true distribution of W. Let $\mathcal{M}$ be a metric space. In our context, we take $\mathcal{M}$ to be the class of all matrices of given dimensions. A matrix-valued statistical functional is a mapping $T : \mathcal{D} \to \mathcal{M}$. When we use the statistic $T(F_n)$ to estimate the parameter $T(F_0)$, we have the following asymptotic result: if $T(F)$ is Fréchet Differentiable at F_0, then

$$T(F_n) = T(F_0) + \mathrm{E}_n T^* + o_p\big(n^{-\frac{1}{2}}\big), \tag{9}$$

where T^* is the influence function of $T(F_n)$ satisfying $\mathrm{E}T^* = 0$. Moreover, the variance matrix of $\mathrm{vec}(T^*)$ has finite entries. The expansion (9) is called the first-order von Mises expansion of $T(F_n)$. By the Central Limit Theorem and Slutzky's theorem, we have

$$\sqrt{n}\big[\mathrm{vec}\big(T(F_n)\big) - \mathrm{vec}\big(T(F_0)\big)\big] \xrightarrow{D} N\big(0, \mathrm{var}\big[\mathrm{vec}\big(T^*\big)\big]\big).$$

The influence function of a statistical functional is defined as follows. Let w be a point in the support of the random vector W. Let δ_w be the Dirac measure at w. Then, the influence function of $T(F_n)$ is defined as the derivative

$$T^* = \frac{d}{d\varepsilon} T\big[(1-\varepsilon)F_0 + \varepsilon\delta_w\big]|_{\varepsilon=0}. \tag{10}$$

The statistics in Section 3, such as the R_n in Theorem 3.3 and $\hat{\mu}_B(\tilde{w})$ in Theorem 3.6, are all special cases of statistical functionals, and their asymptotic distributions can all be derived in this unified fashion.

A.1. Influence functions of Λ_n, Σ_n and $\Sigma_n^{-\frac{1}{2}}$. PROOF. The influence function of Λ_n is simple: recall that $\Lambda_n = \mathrm{E}_n(\mathbb{N}Q\mathbb{N}^T)$. So we have the following expansion:

$$\Lambda_n = \Lambda + \mathrm{E}_n\big(\mathbb{N}Q\mathbb{N}^T - \Lambda\big).$$

The influence function of Λ_n is simply

$$\Lambda^* = \mathbb{N}Q\mathbb{N}^T - \Lambda. \tag{11}$$

We now derive the influence function of Σ_n. Recall that

$$\begin{aligned}\Sigma_n &= \mathrm{var}_n(\mathbb{N}\mathbf{1}/K)\\ &= \mathrm{E}_n\big[\mathbb{N}\big(\mathbf{1}\mathbf{1}^T/K^2\big)\mathbb{N}^T\big] - \mathrm{E}_n(\mathbb{N}\mathbf{1}/K)\mathrm{E}_n\big(\mathbf{1}^T\mathbb{N}^T/K\big).\end{aligned}$$

To calculate the influence function of Σ_n, let $G_\varepsilon = (1-\varepsilon)F_0 + \varepsilon\delta_w$ and

$$\Sigma(\varepsilon) = \int \mathbb{N}\big(\mathbf{1}\mathbf{1}^T/K^2\big)\mathbb{N}^T\, dG_\varepsilon - \int (\mathbb{N}\mathbf{1}/K)\, dG_\varepsilon \int \big(\mathbf{1}^T\mathbb{N}^T/K\big)\, dG_\varepsilon.$$

Because $\frac{d}{d\varepsilon}G_\varepsilon|_{\varepsilon=0} = \delta_w - F_0$, we have

$$\begin{aligned}\left.\frac{d\Sigma(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0} &= \int \mathbb{N}\big(\mathbf{1}\mathbf{1}^T/K^2\big)\mathbb{N}^T\, d(\delta_w - F_0)\\ &\quad - \left[\int (\mathbb{N}\mathbf{1}/K)\, d(\delta_w - F_0)\right]\mathrm{E}\big(\mathbf{1}^T\mathbb{N}^T/K\big)\\ &\quad - \mathrm{E}(\mathbb{N}\mathbf{1}/K)\left[\int \big(\mathbf{1}^T\mathbb{N}^T/K\big)\, d(\delta_w - F_0)\right].\end{aligned}$$

Thus, the influence function for Σ_n is

$$\begin{aligned}\Sigma^* = {} & \mathbb{N}(\mathbf{1}\mathbf{1}^T/K^2)\mathbb{N}^T - \mathrm{E}[\mathbb{N}(\mathbf{1}\mathbf{1}^T/K^2)\mathbb{N}^T] \\ & - [\mathbb{N}\mathbf{1}/K - \mathrm{E}(\mathbb{N}\mathbf{1}/K)]\mathrm{E}(\mathbf{1}^T\mathbb{N}^T/K) \\ & - \mathrm{E}(\mathbb{N}\mathbf{1}/K)[\mathbf{1}^T\mathbb{N}^T/K - \mathrm{E}(\mathbf{1}^T\mathbb{N}^T/K)].\end{aligned} \tag{12}$$

Finally, we derive the influence function of $\Sigma_n^{-\frac{1}{2}}$. Since for all $\varepsilon > 0$, $\Sigma^{-\frac{1}{2}}(\varepsilon)\Sigma(\varepsilon) \times \Sigma^{-\frac{1}{2}}(\varepsilon) = I$, we have

$$(\Sigma^{-\frac{1}{2}})^*\Sigma^{\frac{1}{2}} + \Sigma^{-\frac{1}{2}}\Sigma^*\Sigma^{-\frac{1}{2}} + \Sigma^{\frac{1}{2}}(\Sigma^{-\frac{1}{2}})^* = 0.$$

Hence,

$$(\Sigma^{\frac{1}{2}} \otimes I)\operatorname{vec}[(\Sigma^{-\frac{1}{2}})^*] + (I \otimes \Sigma^{\frac{1}{2}})\operatorname{vec}[(\Sigma^{-\frac{1}{2}})^*] = -(\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}})\operatorname{vec}(\Sigma^*)$$

which implies

$$\begin{aligned}\operatorname{vec}[(\Sigma^{-\frac{1}{2}})^*] &= -(\Sigma^{\frac{1}{2}} \otimes I + I \otimes \Sigma^{\frac{1}{2}})^{-1}(\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}})\operatorname{vec}(\Sigma^*) \\ &= -(\Sigma \otimes \Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}} \otimes \Sigma)^{-1}\operatorname{vec}(\Sigma^*).\end{aligned} \tag{13}$$

□

A.2. Proof of Theorem 3.3.

PROOF. Since $R_n = \Sigma_n^{-\frac{1}{2}}\Lambda_n\Sigma_n^{-\frac{1}{2}}$, we have

$$R(\varepsilon) = \Sigma^{-\frac{1}{2}}(\varepsilon)\Lambda(\varepsilon)\Sigma^{-\frac{1}{2}}(\varepsilon).$$

Differentiate both sides of this equation with respect to ε, and then evaluate the derivatives at $\varepsilon = 0$ to obtain

$$R^* = (\Sigma^{-\frac{1}{2}})^*\Lambda\Sigma^{-\frac{1}{2}} + \Sigma^{-\frac{1}{2}}\Lambda^*\Sigma^{-\frac{1}{2}} + \Sigma^{-\frac{1}{2}}\Lambda(\Sigma^{-\frac{1}{2}})^*.$$

Hence,

$$\begin{aligned}\operatorname{vec}(R^*) &= (\Sigma^{-\frac{1}{2}}\Lambda \otimes I)\operatorname{vec}[(\Sigma^{-\frac{1}{2}})^*] + (\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}})\operatorname{vec}(\Lambda^*) \\ &\quad + (I \otimes \Sigma^{-\frac{1}{2}}\Lambda)\operatorname{vec}[(\Sigma^{-\frac{1}{2}})^*] \\ &= (\Sigma^{-\frac{1}{2}}\Lambda \otimes I + I \otimes \Sigma^{-\frac{1}{2}}\Lambda)\operatorname{vec}[(\Sigma^{-\frac{1}{2}})^*] + (\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}})\operatorname{vec}(\Lambda^*) \\ &= -(\Sigma^{-\frac{1}{2}}\Lambda \otimes I + I \otimes \Sigma^{-\frac{1}{2}}\Lambda)(\Sigma \otimes \Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}} \otimes \Sigma)^{-1}\operatorname{vec}(\Sigma^*) \\ &\quad + (\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}})\operatorname{vec}(\Lambda^*) \\ &= (\Omega_1, \Omega_2)\begin{pmatrix}\operatorname{vec}(\Sigma^*) \\ \operatorname{vec}(\Lambda^*)\end{pmatrix},\end{aligned} \tag{14}$$

where $\Omega_1 = -(\Sigma^{-\frac{1}{2}}\Lambda \otimes I + I \otimes \Sigma^{-\frac{1}{2}}\Lambda)(\Sigma \otimes \Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}} \otimes \Sigma)^{-1}$ and $\Omega_2 = (\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}})$. It follows that

$$\sqrt{n}[\operatorname{vec}(R_n) - \operatorname{vec}(R)] \xrightarrow{D} N(0, \Pi_R),$$

where

$$\Pi_R = (\Omega_1, \Omega_2)\begin{pmatrix}\mathrm{E}[\operatorname{vec}(\Sigma^*)\operatorname{vec}^T(\Sigma^*)] & \mathrm{E}[\operatorname{vec}(\Sigma^*)\operatorname{vec}^T(\Lambda^*)] \\ \mathrm{E}[\operatorname{vec}(\Lambda^*)\operatorname{vec}^T(\Sigma^*)] & \mathrm{E}[\operatorname{vec}(\Lambda^*)\operatorname{vec}^T(\Lambda^*)]\end{pmatrix}\begin{pmatrix}\Omega_1^T \\ \Omega_2^T\end{pmatrix}.$$

We now compute the moments $\mathrm{E}[\mathrm{vec}(\Sigma^*)\,\mathrm{vec}^T(\Sigma^*)]$, $\mathrm{E}[\mathrm{vec}(\Lambda^*)\,\mathrm{vec}^T(\Sigma^*)]$, $\mathrm{E}[\mathrm{vec}(\Lambda^*)\times \mathrm{vec}^T(\Lambda^*)]$. Let $\mathbb{Z} = \mathbb{N}\mathbf{1}/K$. Then, by (12), $\Sigma^* = \mathbb{Z}\mathbb{Z}^T - \mathrm{E}(\mathbb{Z}\mathbb{Z}^T) - [\mathbb{Z} - \mathrm{E}(\mathbb{Z})]\mathrm{E}(\mathbb{Z}^T) - \mathrm{E}(\mathbb{Z})[\mathbb{Z} - \mathrm{E}(\mathbb{Z})]^T$. It follows that $\mathrm{vec}(\Sigma^*) = \mathbb{Z}\otimes\mathbb{Z} - \mathrm{E}(\mathbb{Z}\otimes\mathbb{Z}) - [\mathbb{Z} - \mathrm{E}(\mathbb{Z})]\otimes\mathrm{E}(\mathbb{Z}) - \mathrm{E}(\mathbb{Z})\otimes[\mathbb{Z} - \mathrm{E}(\mathbb{Z})]$. Hence,

$$\begin{aligned}&\mathrm{E}\big[\mathrm{vec}(\Sigma^*)\,\mathrm{vec}^T(\Sigma^*)\big]\\&\quad= \mathrm{E}\big\{\big[\mathbb{Z}\otimes\mathbb{Z} - \mathrm{E}(\mathbb{Z}\otimes\mathbb{Z}) - \big(\mathbb{Z} - \mathrm{E}(\mathbb{Z})\big)\otimes\mathrm{E}(\mathbb{Z}) - \mathrm{E}(\mathbb{Z})\otimes\big(\mathbb{Z} - \mathrm{E}(\mathbb{Z})\big)\big]\\&\qquad\times\big[\mathbb{Z}\otimes\mathbb{Z} - \mathrm{E}(\mathbb{Z}\otimes\mathbb{Z}) - \big(\mathbb{Z} - \mathrm{E}(\mathbb{Z})\big)\otimes\mathrm{E}(\mathbb{Z}) - \mathrm{E}(\mathbb{Z})\otimes\big(\mathbb{Z} - \mathrm{E}(\mathbb{Z})\big)\big]^T\big\}.\end{aligned}$$

Next, by (11), $\Lambda^* = \mathbb{N}Q\mathbb{N}^T - \mathrm{E}(\mathbb{N}Q\mathbb{N}^T)$. It follows that $\mathrm{vec}(\Lambda^*) = (\mathbb{N}\otimes\mathbb{N})\,\mathrm{vec}(Q) - \mathrm{E}[(\mathbb{N}\otimes\mathbb{N})\,\mathrm{vec}(Q)]$. Hence,

$$\begin{aligned}&\mathrm{E}\big[\mathrm{vec}(\Lambda^*)\,\mathrm{vec}^T(\Sigma^*)\big]\\&\quad= \mathrm{E}\big\{\big[(\mathbb{N}\otimes\mathbb{N})\,\mathrm{vec}(Q) - \mathrm{E}\big[(\mathbb{N}\otimes\mathbb{N})\,\mathrm{vec}(Q)\big]\big]\\&\qquad\times\big[\mathbb{Z}\otimes\mathbb{Z} - \mathrm{E}(\mathbb{Z}\otimes\mathbb{Z}) - \big[\mathbb{Z} - \mathrm{E}(\mathbb{Z})\big]\otimes\mathrm{E}(\mathbb{Z}) - \mathrm{E}(\mathbb{Z})\otimes\big[\mathbb{Z} - \mathrm{E}(\mathbb{Z})\big]\big]^T\big\}.\end{aligned}$$

Finally, by (11) again, we have

$$\begin{aligned}\mathrm{E}\big[\mathrm{vec}(\Lambda^*)\,\mathrm{vec}^T(\Lambda^*)\big] &= \mathrm{E}\big\{\big[(\mathbb{N}\otimes\mathbb{N})\,\mathrm{vec}(Q) - \mathrm{E}\big[(\mathbb{N}\otimes\mathbb{N})\,\mathrm{vec}(Q)\big]\big]\\&\qquad\times\big[(\mathbb{N}\otimes\mathbb{N})\,\mathrm{vec}(Q) - \mathrm{E}\big[(\mathbb{N}\otimes\mathbb{N})\,\mathrm{vec}(Q)\big]\big]^T\big\}.\end{aligned}$$

This completes the proof. □

A.3. Proof of Lemma 3.5. Before proving Lemma 3.5, we introduce the following notations. Let $r = K\ell$, and assume matrices R and R_n have eigenvalue decompositions,

$$R = V\Gamma V^T, \qquad R_n = V_n\Gamma_n V_n^T, \tag{15}$$

where $V = (\mathbf{v}_1,\ldots,\mathbf{v}_r)\in\mathbb{R}^{r\times r}$ and $\Gamma = \mathrm{diag}(d_1,\ldots,d_r)\in\mathbb{R}^{r\times r}$, $V_n = (\mathbf{v}_{n,1},\ldots,\mathbf{v}_{n,r})\in\mathbb{R}^{r\times r}$ and $\Gamma_n = \mathrm{diag}(d_{n,1},\ldots,d_{n,r})\in\mathbb{R}^{r\times r}$. Since $\mathbf{b}$ and $\hat{\mathbf{b}}$ are eigenvectors of R and R_n corresponding to their smallest eigenvalues, we have $\mathbf{b} = \mathbf{v}_r$ and $\hat{\mathbf{b}} = \mathbf{v}_{n,r}$. The following lemma is similar but slightly stronger than the assertion 3 of Theorem 3.1.8 of Kollo and von Rosen (2005). The proof is also similar to that given in Kollo and von Rosen (2005) and is omitted.

LEMMA A.1. *Suppose R and R_n have eigenvalue decomposition (15), and satisfy $\sqrt{n}\,\mathrm{vec}(R_n - R)\xrightarrow{D} N(0,\Pi_R)$ as well as Condition 3.4. Then,*

$$\mathbf{v}_{n,i} = \mathbf{v}_i + \big[\mathbf{v}_i^T\otimes V(d_iI - \Gamma)^+V^T\big]\mathrm{E}_n\,\mathrm{vec}(R^*) + o_p\big(n^{-\frac{1}{2}}\big),$$

where $(\cdot)^+$ is the Moore–Penrose inverse.

PROOF OF LEMMA 3.5. Since $\hat{\mathbf{b}}$ and $\mathbf{b}$ are the rth eigenvectors of R_n and R, respectively, the influence function of $\hat{\mathbf{b}}$ is, by Lemma A.1,

$$\mathbf{b}^* = \big[\mathbf{b}^T\otimes V(d_rI - \Gamma)^+V^T\big]\,\mathrm{vec}(R^*) \equiv \Omega_3\,\mathrm{vec}(R^*).$$

Substitute (14) into the right-hand side, to obtain

$$\mathbf{b}^* = (\Omega_3\Omega_1, \Omega_3\Omega_2)\begin{pmatrix}\mathrm{vec}(\Sigma^*)\\ \mathrm{vec}(\Lambda^*)\end{pmatrix}. \tag{16}$$

Hence, $\sqrt{n}(\hat{\mathbf{b}} - \mathbf{b}) \xrightarrow{D} N(\mathbf{0}, \Pi_{\mathbf{b}})$, where $\Pi_{\mathbf{b}} = (\Omega_3\Omega_1, \Omega_3\Omega_2)\Phi(\Omega_3\Omega_1, \Omega_3\Omega_2)^T$. The explicit form of $\Omega_3\Omega_1$ is

$$-\big[\mathbf{b}^T \otimes V(d_r I - \Gamma)^+ V^T\big]\big(\Sigma^{-\frac{1}{2}}\Lambda \otimes I + I \otimes \Sigma^{-\frac{1}{2}}\Lambda\big)\big(\Sigma \otimes \Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}} \otimes \Sigma\big)^{-1},$$

and the explicit form of $\Omega_3\Omega_2$ is

$$\big[\mathbf{b}^T \otimes V(d_r I - \Gamma)^+ V^T\big]\big(\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}}\big).$$

This proves the first part of the lemma.

Now, we turn to the asymptotic distribution of $\hat{\boldsymbol{a}}$. Since $\hat{\boldsymbol{a}} = \Sigma_n^{-\frac{1}{2}}\hat{\mathbf{b}}$, the influence function of $\hat{\boldsymbol{a}}$ is

$$\boldsymbol{a}^* = \big(\Sigma^{-\frac{1}{2}}\big)^*\mathbf{b} + \Sigma^{-\frac{1}{2}}\mathbf{b}^* = \big(\mathbf{b}^T \otimes I\big)\operatorname{vec}\big[\big(\Sigma^{-\frac{1}{2}}\big)^*\big] + \Sigma^{-\frac{1}{2}}\mathbf{b}^*.$$

Substitute (13) and (16) into the right-hand side to obtain

$$\begin{aligned}
\boldsymbol{a}^* &= -\big(\mathbf{b}^T \otimes I\big)\big(\Sigma \otimes \Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}} \otimes \Sigma\big)^{-1}\operatorname{vec}\big(\Sigma^*\big) + \Sigma^{-\frac{1}{2}}\Omega_3\Omega_1\operatorname{vec}\big(\Sigma^*\big) \\
&\quad + \Sigma^{-\frac{1}{2}}\Omega_3\Omega_2\operatorname{vec}\big(\Lambda^*\big) \\
&= \big[-\big(\mathbf{b}^T \otimes I\big)\big(\Sigma \otimes \Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}} \otimes \Sigma\big)^{-1} + \Sigma^{-\frac{1}{2}}\Omega_3\Omega_1\big]\operatorname{vec}\big(\Sigma^*\big) \\
&\quad + \Sigma^{-\frac{1}{2}}\Omega_3\Omega_2\operatorname{vec}\big(\Lambda^*\big) \\
&\equiv \Omega_4\operatorname{vec}\big(\Sigma^*\big) + \Omega_5\operatorname{vec}\big(\Lambda^*\big).
\end{aligned}$$

It follows that $\sqrt{n}(\hat{\boldsymbol{a}} - \boldsymbol{a}) \xrightarrow{D} N(\mathbf{0}, \Pi_a)$, where $\Pi_a = (\Omega_4, \Omega_5)\Phi(\Omega_4, \Omega_5)^T$, with Φ being given in Theorem 3.3. We now derive the explicit forms of Ω_4 and Ω_5. By definition,

$$\begin{aligned}
\Omega_4 &= -\big(\mathbf{b}^T \otimes I\big)\big(\Sigma \otimes \Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}} \otimes \Sigma\big)^{-1} + \Sigma^{-\frac{1}{2}}\Omega_3\Omega_1 \\
&= -\big(\mathbf{b}^T \otimes I\big)\big(\Sigma \otimes \Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}} \otimes \Sigma\big)^{-1} \\
&\quad - \Sigma^{-\frac{1}{2}}\big[\mathbf{b}^T \otimes V(d_r I - \Gamma)^+ V^T\big]\big(\Sigma^{-\frac{1}{2}}\Lambda \otimes I + I \otimes \Sigma^{-\frac{1}{2}}\Lambda\big)\big(\Sigma \otimes \Sigma^{\frac{1}{2}} + \Sigma^{\frac{1}{2}} \otimes \Sigma\big)^{-1}.
\end{aligned}$$

Also,

$$\Omega_5 = \Sigma^{-\frac{1}{2}}\big[\mathbf{b}^T \otimes V(d_r I - \Gamma)^+ V^T\big]\big(\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}}\big).$$

The proof is completed. □

A.4. Proof of Theorem 3.6. PROOF. Recall that $\hat{\mu}_B(\tilde{w}) = \hat{\boldsymbol{a}}^T\mathbb{N}(\tilde{w})\mathbf{1}/K$, where $\mathbb{N}(\tilde{w})$ is the nonrandom matrix

$$\operatorname{diag}\big(\mathbf{N}_1(\tilde{w}_1), \ldots, \mathbf{N}_K(\tilde{w}_K)\big).$$

Hence, $\sqrt{n}[\hat{\mu}_B(\tilde{w}) - \mu_B(\tilde{w})] = (\hat{\boldsymbol{a}} - \boldsymbol{a})^T\mathbb{N}(\tilde{w})\mathbf{1}/K$. By the asymptotic distribution of $\hat{\boldsymbol{a}}$ in Lemma 3.5, we have

$$\sqrt{n}\big[\hat{\mu}_B(\tilde{w}) - \mu_B(\tilde{w})\big] \xrightarrow{D} N\big(0, \sigma_\mu^2\big), \tag{17}$$

where $\sigma_\mu^2 = \mathbf{1}^T\mathbb{N}^T(\tilde{w})(M_{a1}, M_{a2})\Phi(M_{a1}, M_{a2})^T\mathbb{N}(\tilde{w})\mathbf{1}/K^2$, $M_{a1} = \Omega_4$, and $M_{a2} = \Omega_5$. The vector $\mathbb{N}(\tilde{w})\mathbf{1}$ is simply $\mathbf{N}(\tilde{w})$, defined in Theorem 3.6. □

Acknowledgments. The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding. Zhong's research was supported by U.S. National Science Foundation under grants DMS-1440037, DMS-1440038, and DMS-1438957 and by U.S. National Institute of Health under grants R01GM122080 and R01GM113242. Li's research was supported by U.S. National Science Foundation under grant DMS-1713078.

SUPPLEMENTARY MATERIAL

Supplement A (DOI: 10.1214/21-AOAS1537SUPPA; .pdf). We discuss the implementation issues of the B-scaling method and report results from additional simulation studies.

Supplement B (DOI: 10.1214/21-AOAS1537SUPPB; .zip). We implemented the B-scaling algorithm using R. An example is available on GitHub.

REFERENCES

AMIN, V., HARRIS, R. A., ONUCHIC, V., JACKSON, A. R., CHARNECKI, T., PAITHANKAR, S., SUBRAMANIAN, S. L., RIEHLE, K., COARFA, C. et al. (2015). Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. *Nat. Commun.* **6** 1–10.

BERNSTEIN, B. E., STAMATOYANNOPOULOS, J. A., COSTELLO, J. F., REN, B., MILOSAVLJEVIC, A., MEISSNER, A., KELLIS, M., MARRA, M. A., BEAUDET, A. L. et al. (2010). The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **28** 1045–1048.

BIRD, A. (2007). Perceptions of epigenetics. *Nature* **447** 396–398. https://doi.org/10.1038/nature05913

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*: *A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. CRC Press/CRC, Boca Raton, FL. MR2243417 https://doi.org/10.1201/9781420010138

COX, T. F. and COX, M. A. A. (2000). *Multidimensional Scaling*. Chapman and Hall/CRC.

EGGER, G., LIANG, G., APARICIO, A. and JONES, P. A. (2004). Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **429** 457.

ERNST, J. and KELLIS, M. (2012). Chromhmm: Automating chromatin-state discovery and characterization. *Nat. Methods* **9** 215–216.

ESTELLER, M. (2008). Epigenetics in cancer. *N. Engl. J. Med.* **358** 1148–1159.

FEINBERG, A. P., KOLDOBSKIY, M. A. and GÖNDÖR, A. (2016). Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat. Rev. Genet.* **17** 284–299.

FERNHOLZ, L. T. (2012). *Von Mises Calculus for Statistical Functionals*. *Lecture Notes in Statistics* **19**. Springer Science & Business Media, New York.

FULLER, W. A. (2009). *Measurement Error Models*. *Wiley Series in Probability and Statistics*, *Vol.* 305. Wiley, Hoboken, NJ.

GOMEZ-CABRERO, D., ABUGESSAISA, I., MAIER, D., TESCHENDORFF, A., MERKENSCHLAGER, M., GISEL, A., BALLESTAR, E., BONGCAM-RUDLOFF, E., CONESA, A. et al. (2014). Data integration in the era of omics: Current and future challenges. *BMC Syst. Biol.* **8** I1.

HALL, D. L. and LLINAS, J. (1997). An introduction to multisensor data fusion. *Proc. IEEE* **85** 6–23.

HALL, D. L. and MCMULLEN, S. A. (2004). *Mathematical Techniques in Multisensor Data Fusion*. Artech House.

HAMID, J. S., HU, P., ROSLIN, N. M., LING, V., GREENWOOD, C. M. and BEYENE, J. (2009). Data integration in genetics and genomics: Methods and challenges. In *Human Genomics and Proteomics*: *HGP*.

HE, X., SHEN, L. and SHEN, Z. (2001). A data-adaptive knot selection scheme for fitting splines. *IEEE Signal Process. Lett.* **8** 137–139.

KHALEGHI, B., KHAMIS, A., KARRAY, F. O. and RAZAVI, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Inf. Fusion* **14** 28–44.

KOLLO, T. and VON ROSEN, D. (2005). *Advanced Multivariate Statistics with Matrices*. *Mathematics and Its Applications* (*New York*) **579**. Springer, Dordrecht. MR2162145 https://doi.org/10.1007/1-4020-3419-9

KRUSKAL, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29** 1–27. MR0169712 https://doi.org/10.1007/BF02289565

KUNDAJE, A., MEULEMAN, W., ERNST, J., BILENKY, M., BERNSTEIN, B. E., COSTELLO, J. F., ECKER, J. R., HIRST, M., MEISSNER, A. et al. (2015). February. Integrative analysis of 111 reference human epigenomes. *Nature* **518** 317–330.

LEE, L., WANG, K., LI, G., XIE, Z., WANG, Y., XU, J., SUN, S., POCALYKO, D., BHAK, J. et al. (2011). Liverome: A curated database of liver cancer-related gene signatures with self-contained context information. *BMC Genomics* **12** S3.

LETUNIC, I., COPLEY, R. R., SCHMIDT, S., CICCARELLI, F. D., DOERKS, T., SCHULTZ, J., PONTING, C. P. and BORK, P. (2004). Smart 4.0: Towards genomic data integration. *Nucleic Acids Res*. **32** D142–D144.

LI, B. (2018). Linear operator-based statistical analysis: A useful paradigm for big data. *Canad. J. Statist*. **46** 79–103. MR3767167 https://doi.org/10.1002/cjs.11329

LI, B. and WANG, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc*. **102** 997–1008. MR2354409 https://doi.org/10.1198/016214507000000536

LIU, Y. SUN, X. ZHONG, W. and LI, B. (2022). Supplement to "B-scaling: A novel nonparametric data fusion method." https://doi.org/10.1214/21-AOAS1537SUPPA, https://doi.org/10.1214/21-AOAS1537SUPPB

MENG, C., ZELEZNIK, O. A., THALLINGER, G. G., KUSTER, B., GHOLAMI, A. M. and CULHANE, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform*. **17** 628–641.

PORTELA, A. and ESTELLER, M. (2010). Epigenetic modifications and human disease. *Nat. Biotechnol*. **28** 1057–1068.

RAY, B., LIU, W. and FENYÖ, D. (2017). Adaptive multiview nonnegative matrix factorization algorithm for integration of multimodal biomedical data. *Cancer Inform*. **16**.

REUTER, J. A., SPACEK, D. V. and SNYDER, M. P. (2015). High-throughput sequencing technologies. *Mol. Cell* **58** 586–597.

RITCHIE, M. D., HOLZINGER, E. R., LI, R., PENDERGRASS, S. A. and KIM, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet*. **16** 85.

ROMANOSKI, C. E., GLASS, C. K., STUNNENBERG, H. G., WILSON, L. and ALMOUZNI, G. (2015). February. Epigenomics: Roadmap for regulation. *Nature* **518** 314–316.

SHEAFFER, K. L., KIM, R., AOKI, R., ELLIOTT, E. N., SCHUG, J., BURGER, L., SCHÜBELER, D. and KAESTNER, K. H. (2014). DNA methylation is required for the control of stem cell differentiation in the small intestine. *Genes Dev*. **28** 652–664.

THE ENCODE PROJECT CONSORTIUM (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57.

TORGERSON, W. S. (1952). Multidimensional scaling. I. Theory and method. *Psychometrika* **17** 401–419. MR0054219 https://doi.org/10.1007/BF02288916

WALTZ, E., LLINAS, J. et al. (1990). *Multisensor Data Fusion* **685**. Artech House, Boston, MA.

YUAN, Y., CHEN, N. and ZHOU, S. (2013). Adaptive B-spline knot selection using multi-resolution basis set. *IIE Trans*. **45** 1263–1277.

YUAN, G.-C., MA, P., ZHONG, W. and LIU, J. S. (2006). Statistical assessment of the global regulatory role of histone acetylation in saccharomyces cerevisiae. *Genome Biol*. **7** R70.

ZANG, C., WANG, T., DENG, K., LI, B., HU, S., QIN, Q., XIAO, T., ZHANG, S., MEYER, C. A. et al. (2016). High-dimensional genomic data bias correction and data integration using MANCIE. *Nat. Commun*. **7** 1–8.

ZHANG, S., LI, Q., LIU, J. and ZHOU, X. J. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics* **27** i401–i409.

ZHANG, S., LIU, C.-C., LI, W., SHEN, H., LAIRD, P. W. and ZHOU, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res*. **40** 9379–9391.