Preprint (2022), **0**, 0, *pp.* 1–26
doi:

A structured multivariate approach for removal of latent batch effects

RONGQIAN ZHANG

Department of Statistical Sciences, University of Toronto

LINDSAY D. OLIVER

The Centre for Addiction and Mental Health

ARISTOTLE N. VOINESKOS

Department of Psychiatry, University of Toronto & the Centre for Addiction and Mental Health

JUN YOUNG PARK*

Department of Statistical Sciences and Department of Psychology, University of Toronto

Summary

Combining data collected from multiple studies is becoming common and is advantageous to researchers to increase the reproducibility of scientific discoveries. However, at the same time, unwanted "batch effects" are commonly observed across neuroimaging data collected from multiple study sites or scanners, rendering difficulties in combining such data to obtain reliable findings. While methods for handling such unwanted variations have been proposed recently, most of them use univariate approaches which would be too simple to capture all sources of batch effects which could be represented by the batch-specific latent patterns. In this paper, we propose a novel

*To whom correspondence should be addressed.

multivariate harmonization method, called UNIFAC harmonization, for estimating and removing both explicit and latent batch effects. Our approach is based on the simultaneous dimension reduction and factorization of interlinked matrices through a penalized objective, which provides a new direction in neuroimaging research for harmonizing multivariate features across batches. Using the Social Processes Initiative in Neurobiology of the Schizophrenia (SPINS) dataset and extensive simulation studies, we show that UNIFAC harmonization performed better than the existing methods in entirely removing batch effects as well as retaining associations of interest to increase statistical power. The proposed method is publicly available as a R package.

Key words: Batch effects; Covariance heterogeneity; Dimension reduction; Genomics; Neuroimaging; UNIFAC harmonization

1. Introduction

1.1 *Overview*

It is increasingly common in neuroimaging and genomics to combine data collected from multiple studies to increase the power and the reproducibility of scientific discoveries. However, combining such data comes with unwanted variations, termed batch effects, that must be removed for successful data integration. For example, in neuroimaging studies, study sites often use scanners with different optimization protocols (Fortin *and others*, 2017, 2018; Yu *and others*, 2018). Similarly, genome-wide RNA expression studies involve different sample preparation and sequencing methods (Johnson *and others*, 2007). These heterogeneous data preparation pipelines can lead to batch effects and incorrect conclusions. In this paper, we will use scanner effects to denote batch effects in neuroimaging data.

In the last two decades, there have been numerous efforts in statistics to capture and remove these unwanted variations and increase the signal-to-noise ratio. It is exemplified by the ComBat

method, which has been applied to many data types in genomics (microarray and RNA-Seq) (Johnson *and others*, 2007; Zhang *and others*, 2020) and neuroimaging (functional, structural, and diffusion magnetic resonance imaging) (Fortin *and others*, 2017, 2018; Yu *and others*, 2018). However, most of the methods, including ComBat are univariate approaches that would be limited to capturing all sources of batch effects which could be represented by the *batch-specific latent patterns*. (Chen *and others*, 2022)

We hypothesize that a principled and fully multivariate approach in this paper can further improve the data quality and reproducibility of findings by capturing latent patterns of batch effects, and this novel batch-correction method is based on the dimension reduction and factorization of interlinked matrices.

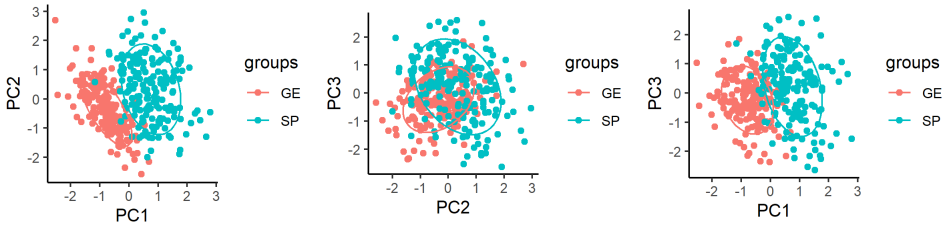
1.2 *Existing methods*

ComBat (Johnson *and others*, 2007) is an empirical Bayes method, where batch effects are characterized by additive batch effects (locations) and multiplicative batch effects (scales). ComBat has been shown to be more robust to outliers in the case of small within-batch sample sizes (Johnson *and others*, 2007; Yu *and others*, 2018). However, ComBat is limited by the assumption that additive batch effects can be explained by only an intercept for each scanner and feature. This oversimplified assumption may ignore the unknown or partially known latent pattern batch effects and may include unwanted batch-specific latent patterns in the harmonized data, further resulting in the potential loss of power. CovBat (Chen *and others*, 2022), a recent multivariate batch-correction extending ComBat, takes the covariance of multivariate features into consideration for these latent batch effects. It applies ComBat twice: first to the original data, then to the principal components from the residual matrix. CovBat assumes the principal components from the residuals follow a ComBat model, which might not be a sufficient assumption to characterize all existing batch-specific patterns.

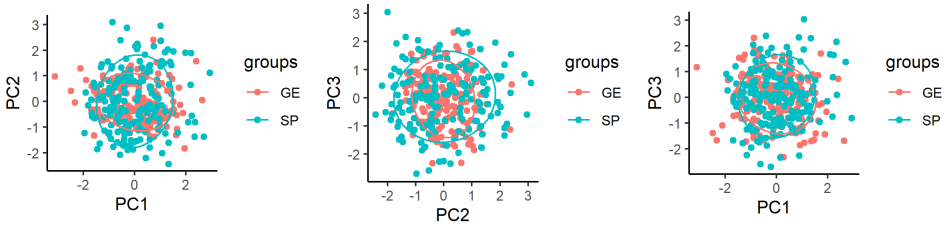
SVA (Surrogate Variable Analysis) is another method that was originally developed for genomic studies (Leek and Storey, 2007) then adapted to neuroimaging studies (Fortin *and others*, 2016). It includes latent factors of unwanted variation as surrogate variables, which are not associated with the biological covariates of interest. Instead of using explicit variables to denote batch effects, SVA identifies and estimates unwanted variations, possibly including batch and other artifacts, through permutation testing then removes them as surrogate variables. RAVEL (Fortin *and others*, 2016) is a batch effect correction method for neuroimaging data inspired by RUV (Gagnon-Bartsch *and others*, 2013). It estimates and removes unwanted variation factors by using negative controls, which are features (e.g. genes, voxels, etc.) that are known a priori to be unassociated with the variables of interest (Fortin *and others*, 2017, 2016). This method applies singular value decomposition (SVD) to obtain latent factors of unwanted variations in the control regions then removes the latent factors and corresponding effects in the test regions. Although many studies show the discussed methods are applicable to different data types, some are built on univariate linear regression framework (i.e., ComBat, CovBat, RAVEL) (Fortin *and others*, 2017; Chen *and others*, 2022). In addition, methods such as SVA (Leek and Storey, 2007) and RAVEL (Fortin *and others*, 2016) are confined to capturing all unwanted variations or artifacts, not only batch-specific variations, and require external data for formulations.

1.3 *Motivating example*

The Social Processes Initiative in Neurobiology of the Schizophrenia (SPINS) is a large multi-site, multi-scanner study examining social cognition in schizophrenia, with data for T1- weighted, diffusion-weighted, and resting-state functional magnetic resonance imaging (MRI) scans from both individuals with schizophrenia and healthy controls as well as demographic information. More detailed information about the study and data will be described in Section 3.1. Study subjects were recruited from 3 study sites. Initially, subjects were scanned by General Electric



(i) The first 3 principal components (PCs) of unharmonized data after regressing out potential covariates



(ii) The first 3 principal components (PCs) of ComBat harmonized data after regressing out potential covariates

Fig. 1: Visualization of scanner effects in the SPINS data before and after the ComBat harmonization, colored by scanner (GE: General Electric, SP: Siemens Prisma).

(750w Discovery or Signa) or Siemens Tim Trio based on the location, but during the study Siemens Prisma was uniformly used to scan the brain. By applying principal component analysis (PCA) to the original fractional anisotropy (FA) data from the diffuson tensor imaging (DTI), we observed that the most variations are clearly explained by the scanner information (General Electric vs. Siemens Prisma) (Figure 1(i)). We then applied ComBat and extracted top 3 principal components from the harmonized data to see if the data reveals any remaining scanner-specific variations (Figure 1(ii)). Despite an evidence of higher data quality by ComBat than the original data, the heteroscedasticity of ellipses in top principal components indicates there are still unremoved latent patterns specific to scanner information. This motivates a need for a new batch-correction method that successfully removes *hidden* scanner-specific patterns not captured by the simple location-scale adjustment.

1.4 *Our contribution*

We propose a novel method, called *UNIFAC harmonization*, for estimating and removing both explicit and latent batch effects. It extends the work of Park and Lock (2020), a dimension reduction method primarily motivated by integrating multiple data types. While their work allows flexible formulation of wanted and unwanted variations into the low-rank approximations, its formulations and applications to the batch-correction context as well as comparisons to existing methods have not been explored yet. We show that the proposed multivariate method simultaneously identifies and corrects not only explicit additive batch effects (locations) and multiplicative batch effects (scales), but also latent batch effects. The latent variable formulation gives a clear understanding of the method in terms of harmonizing *covariances* across batches.

The rest of the papers are organized as follows. Section 2 describes our proposed method, UNIFAC harmonization, and describes the relationship between UNIFAC harmonization and ComBat. At the same time, we reformulate the existing "joint and individual" factorization methods into the harmonization context, which provides an intuitive explanation on covariance heterogeneities. In Section 3, we harmonize the SPINS data with UNIFAC harmonization and compare it to other harmonization methods using a comprehensive evaluation framework. Section 4 conducts simulations to evaluate performances in terms of Type 1 error rate and statistical power. We conclude with some points of discussions in Section 5.

2. Methods

2.1 *Notation and setup*

Let $\mathbf{Y}$ be a $p \times n$ data matrix of p features for n subjects. Then, consider $\{\mathbf{Y}_j : p \times n_j | j = 1, ..., J\}$ be a partition of $\mathbf{Y}$, where J is the number of batches and the jth batch has n_j subjects (so that $n = \sum_{j=1}^{J} n_j$). The matrices can be concatenated to form a matrix $\mathbf{Y} = [\mathbf{Y}_1; \mathbf{Y}_2; \ldots; \mathbf{Y}_J]$. We

will use this notation for a general $p \times n$ matrix throughout this article.

We first characterize additive and multiplicative batch effects by using a factorizable multivariate model. We assume that the data matrix $\mathbf{Y}$ is decomposed into

$$\mathbf{Y} = \mathbf{R} + [\mathbf{I}_1; \ldots; \mathbf{I}_J] + [\delta_1 \mathbf{E}_1; \ldots; \delta_J \mathbf{E}_J], \tag{2.1}$$

where $\mathbf{R}$ is a $p \times n$ low-rank row-shared structure containing information shared across all batches horizontally, which should be retained after harmonization. Each $\mathbf{I}_j$ is a $p \times n_j$ low-rank latent patterns containing batch effects shared only in each batch j, which needs to be removed after harmonization. $\mathbf{E}_j$ is a full-rank noise matrix with a unit variance, and δ_j^2 characterizes the noise variance for j which is assumed to be heterogeneous.

2.2 *UNIFAC harmonization*

Our approach is summarized by (i) removing batch and feature-specific means first, (ii) standardizing the data matrix to have homogeneous variance, (iii) decomposing it into batch specific and batch independent factors, and (iv) reconstructing harmonized data.

Steps (i) and (ii) are achieved through the preprocessing step. We first row-center each $\mathbf{Y}_j$ to have zero mean and scale each row of $\mathbf{Y}$ to have a unit variance for $\mathbf{Y}$. This step guarantees that each $\mathbf{Y}_j$ has zero mean and consists of low-rank signals plus Gaussian noise with variance δ_j^2. From here, we scale each data matrix $\mathbf{Y}_j$ by $\hat{\delta}_j$. Following Park and Lock (2020) and Lock *and others* (2022), we estimate $\hat{\delta}_j$ using random matrix theory, specifically, by the median of the singular values of $\mathbf{Y}_j$ divided by the square root of the median of the Marcenko-Pastur distribution (Gavish and Donoho, 2017). We use $\mathbf{Y}_j^\star$ to denote the result from Step (i) and (ii). This processing will not affect the estimation of the low-rank structure of the batch independent components and latent batch effects because they are reserved for Step (iii).

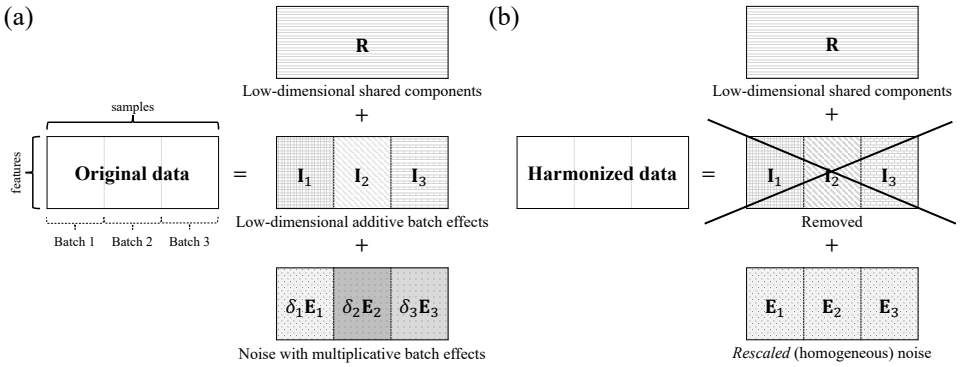


Fig. 2: Overview of the proposed model, using 3-batch data structure

In Step (iii), provided that $\hat{\delta}_j \approx \delta_j$,we first note that $\mathbf{Y}^\star$ is represented by

$$\mathbf{Y}^\star = \mathbf{R}^\star + \mathbf{I}^\star + \mathbf{E} \tag{2.2}$$

where $\mathbf{R}^\star$ is a row-shared structure containing information shared across all batches, $\mathbf{I}^\star = [\mathbf{I}_1^\star; \ldots; \mathbf{I}_J^\star]$ are individual structures containing information shared only in each individual batch.

Due to low-rank approximations, $\mathbf{R}^\star$ and $\mathbf{I}^\star$ can be written as a product of row loadings and column scores. The row-shared structures $\mathbf{R}^\star$ have common loadings across all batches, and the scores can represent important biological covariates in a study. The individual structure $\mathbf{I}^\star$ have different loadings and scores for different batches. The scores for $\mathbf{I}^\star$ can represent non-biological covariates due to different scanners or sites in a study.

We use a Figure 2 to illustrate the proposed factorized forms. In this three-batch structures, $\mathbf{R}$ explains substantial variations shared by all batches. Thus, the low-rank shared effects $\mathbf{R}$ will be retained in the harmonization. The loadings of the individual structures $\mathbf{I}_j$, include batch effects that explain variations in only j-th batch. So, these low-rank individual effects $\mathbf{I}_j$ $(j = 1 \ldots, 3)$ will be artificial (non-biological) variations that should be removed in the harmonization. The noise structure $\delta_j \mathbf{E}_j$ with multiplicative batch effect has been scaled by $\hat{\delta}_j$ to be a homogeneous noise in Step (ii).

For estimation, we use a penalized objective function extending the nuclear norm penalization

for a single matrix. Based on the model (2.2), $\widehat{\mathbf{R}}^{\star}$ and $\widehat{\mathbf{I}}^{\star}$ are obtained by

$$\{\widehat{\mathbf{R}}^{\star}, \widehat{\mathbf{I}}^{\star}\} = \underset{\{\mathbf{R}^{\star}, \mathbf{I}^{\star}\}}{\arg\min} \left\{ ||\mathbf{Y}^{\star} - \mathbf{R}^{\star} - \mathbf{I}^{\star}||_F^2 + \lambda ||\mathbf{R}^{\star}||_* + \sum_{j=1}^{J} \lambda_j ||\mathbf{I}_j^{\star}||_* \right\}, \tag{2.3}$$

where $||\cdot||_F$ and $||\cdot||_*$ are the Frobenious norm and the nuclear norm, respectively. Its objective function is an extension of softImpute that extracts low-rank signals from a single data matrix. The nuclear norm penalties in the object ensure that the resulting $\widehat{\mathbf{R}}^{\star}, \widehat{\mathbf{I}}^{\star}$ are low-rank, thus it achieves simultaneous dimension reduction and estimation. Due to the convexity of the objective, the block-wise coordinate descent algorithm can be applied to obtain $\widehat{\mathbf{R}}^{\star}$ and $\widehat{\mathbf{I}}_j^{\star}$ (Park and Lock, 2020). Because $\mathbf{Y}^{\star}$ is a zero mean matrix with low-rank signals and independent Gaussian noise with the unit variance, we use the recommended values from Park and Lock (2020) by setting $\lambda = \sqrt{p} + \sqrt{n}$ and $\lambda_j = \sqrt{p} + \sqrt{n_j}$ as probabilistic upper bounds of the largest singular values of $\mathbf{E}$ and $\mathbf{E}_j$, respectively. Park and Lock (2020) and Lock *and others* (2022) showed that these tuning parameters meet necessary conditions for identifiability.

In Step (iv), since we hope to keep the shared effects to the original scale and remove multiplicative batch effects only for noise structure, and we standardize all the components including $\mathbf{R}$ in the Step (ii), we need to scale $\mathbf{R}^{\star}$ back as $\hat{\delta}_j \widehat{\mathbf{R}}_j^{\star}$. We also hope to keep the noise structures to the original scale but with homogeneous variance, so we need to scale $\mathbf{E}$ back as $\hat{\delta}\widehat{\mathbf{E}}$, where $\hat{\delta}^2 = (\sum_{j=1}^{J} n_j \hat{\delta}_j^2)/(\sum_{j=1}^{J} n_j)$. Therefore, the final UNIFAC-harmonized results are defined as $\mathbf{Y}^{UNIFAC} = \left[\widetilde{\mathbf{Y}}_1, \ldots, \widetilde{\mathbf{Y}}_J\right]$, where

$$\widetilde{\mathbf{Y}}_j = \underbrace{\hat{\delta}_j \widehat{\mathbf{R}}_j^{\star}}_{\text{original-scale shared effects}} + \underbrace{\hat{\delta} \times (\mathbf{Y}_j^{\star} - \widehat{\mathbf{R}}_j^{\star} - \widehat{\mathbf{I}}_j^{\star})}_{\text{rescaled noise}} \tag{2.4}$$

.

2.3 *Notes on using covariates*

Adjusting for baseline covariates is straightforward by regressing out covariate effects first, obtaining harmonized "residuals", and adding covariate effects back. However, if one's interest is conducting association testing with covariates, including covariates of interests in UNIFAC harmonization may lead to inflated false positive findings. It is because our objective (2.3) does not enforce scores of $\widehat{\mathbf{I}}^{\star}$ to be independent with covariates of interest. Therefore, we suggest not including variables of interest when applying UNIFAC harmonization and use it for testing. In practice, we found that even not including any covariates does not result in a noticeable difference. It is because the covariate effects are actually a low-rank (with rank equal to the number of covariates) batch-independent patterns and are captured by $\mathbf{R}$ (in high signal-to-noise ratio (SNR)) or by $\mathbf{E}$ (in low SNR), provided that covariates are indepedent to batch information.

However, we note that ComBat does not suffer from potential inflated false positives provided that the variables of interest and the indicator variable for additive batch effect are independent (Nygaard *and others*, 2016).

2.4 *Relationship between UNIFAC harmonization and ComBat*

For feature $i = 1, \ldots, p$ and subject $k = 1, \ldots, n_j$, ComBat uses $y_{ijk} = \mathbf{x}_k^T \boldsymbol{\beta}_i + \gamma_{ij} + \phi_{ij}\epsilon_{ijk}$ to specify additive and multiplicative batch effects. In ComBat, the covariate effects are the only explicit shared information preserved in harmonized data. Therefore, although ϵ_{ijk} may contain (i) shared information *not* explained by linear covariate effects including non-linear covariate effects or unobserved covariate effects and (ii) batch-specific variations not captured by batch-specific means, ComBat does not distinguish these and leave (ii) unremoved. While regressing out covariate effects ($\mathbf{x}_k^T \boldsymbol{\beta}_i$) and additive batch effects (γ_{ij}) remains the same in UNIFAC harmonization in the preprocessing step, it further decomposes ϵ_{ijk} in a more interpretable way. Also, while ϕ_{ij} in ComBat and δ_j in UNIFAC harmonization both account for heterogeneous noise variances,

the "noise" terms are treated differently. UNIFAC harmonization models heterogeneity in "white noise", but ComBat models heterogeneity in residuals not explained by covariates.

2.5 *Reformulation of other matrix factorization methods in data harmonization context*

The key idea of UNIFAC is in line with other "joint and individual" factorization of interlinked matrices, which were primarily developed for decomposing *multiple-omics* data collected in *a single cohort* into a set of low-rank modules. JIVE (Lock *and others*, 2013) is a classic method in multi-omic studies that uses permutation or BIC to choose ranks in both shared and individual structures. Other methods include AJIVE (Feng *and others*, 2018), SLIDE (Gaynanova and Li, 2019) and others, which expand JIVE on interpretability, rank selection (e.g. bi-cross validation), and allowance of "partially-shared" modules.

These can be adapted for harmonizing data collected from multiple studies, as transposing vertically stacked matrices allows us to formulate *multiple cohorts* (batches) for *a single data type*. In their harmonization, instead of centering by feature for $\mathbf{Y}$ in data integration, we remove batch and feature-specific means. Following standard practices, we then scale each data matrix by the Frobenius norm of the data matrix to have homogeneous noise. For rank selection, we use prespecified rank for AJIVE (e.g. scree plots) or the estimated rank from SLIDE (e.g. bi-cross validation). Finally, we rescale shared components and noise terms analogous to UNIFAC and construct harmonized data.

UNIFAC harmonization has some advantages over other methods in the data harmonization context. First, tuning parameters are interpretable in terms of white noise, which aligns well with the harmonization context. Second, it was shown that the proposed objective is superior to other methods in various signal-to-noise scenarios (Park and Lock, 2020). Third, the tuning parameter selection is determined by random matrix theory so it is computationally more efficient than other methods. In Section 3 and 4, we provide a comprehensive evaluation of UNIFAC in comparison

to other methods constructed in a harmonization context.

3. Real Data Analysis

3.1 *Data preparation*

We used diffusion tensor imaging (DTI) data from the Social Processes Initiative in Neurobiology of the Schizophrenia (SPINS) study to empirically evaluate UNIFAC harmonization's performance. Study subjects consisted of 256 individuals with schizophrenia spectrum disorders (SSDs) and 175 controls. Participants with SSDs met DSM-5 diagnostic criteria for schizophrenia, schizoaffective disorder, schizophreniform disorder, delusional disorder, or psychotic disorder not otherwise specified, assessed using the Structured Clinical Interview for DSM (SCID-IV-TR), and had no change in antipsychotic medication or decrement in functioning/support level in the 30 days prior to enrollment. Controls did not have a current or past Axis I psychiatric disorder, excepting adjustment disorder, phobic disorder, and past major depressive disorder (over two years prior; presently unmedicated), or a first degree relative with a history of psychotic mental disorder. Additional exclusion criteria included a history of head trauma resulting in unconsciousness, a substance use disorder (confirmed by urine toxicology screening), intellectual disability, debilitating or unstable medical illness, or other neurological diseases. Participants also had normal or corrected-to-normal vision.

Subjects were 18–55 years old, and 268 of the participants were males (163 females). Participants' white matter tracts were reconstructed using deterministic unscented Kalman Filter (UKF) tractography (Malcolm *and others*, 2010) in 3D Slicer (`https://github.com/SlicerDMRI`). The ORG (O'Donnell Research Group) white matter atlas (Zhang and others, 2018) was used to parcellate fibers into anatomical tracts. Metrics were included from 56 deep white matter fiber tracts from the association, cerebellar, commissural and projection tracts (the cortico-ponto-cerebellar tract was excluded due to parcellation issues), and 16 superficial tract categories according to the

brain lobes they connect, resulting in $p = 72$ features. Mean fractional anisotropy (FA) values were calculated along each tract. FA measures the degree to which diffusion of water molecules is restricted by microstructural elements such as cell bodies, axons, myelin, and other constituents of cytoskeleton (Beaulieu, 2002). Visual quality control was performed after initial tractography, registration to the ORG atlas, and tract creation. Data from seven participants were excluded on the basis of missing or poor tractography for > 15 tracts across the whole brain.

The scans were acquired at three different imaging sites, including Center for Addiction and Mental Health (CAMH), Maryland Psychiatric Research Center (MPRC), and Zucker Hillside Hospital (ZHH). The MRI scanner used at CAMH and ZHH was a General Electric (750w Discovery and Signa respectively), and the MRI scanner at MPRC was a Siemens Tim Trio. However, during the middle of the study, all study sites switched to Siemens Prisma for data collection. Since the number of samples from Siemens Tim Trio are small, we dropped ST and used two scanner types (GE and SP) in our analysis. Participants without DTI data were excluded from the study, leaving us with a final dataset of $n = 351$ subjects.

Table 1: Description of study subjects in the SPINS dataset

Scanner	# subjects	# female(%)	Age (range)	# disease(%)
General Electric (GE)	172	67(39)	[18,55]	111(65)
Siemens Prisma (SP)	179	71(40)	[18,55]	98(55)

3.2 *Results*

We fitted the UNIFAC harmonization, ComBat, CovBat, Adjusted Residuals as well as AJIVE and SLIDE modified in the harmonization context described in Section 2.3. Adjusted Residuals harmonization only removes additive batch effects by adjusting for biological covariates. We used age, age^2, gender, disease status as covariates, in which we later used an interaction between age and disease status to evaluate statistical power of the harmonization methods.

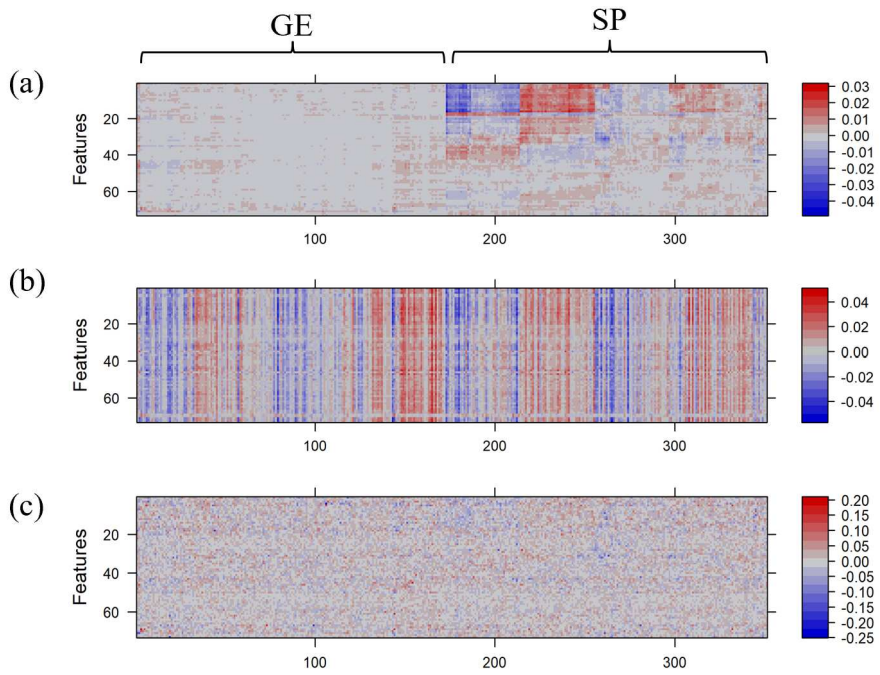


Fig. 3: Heatmaps of the 3 components of UNIFAC harmonization (a) Individual latent pattern ($\widehat{\mathbf{I}}$). (b) Row-shared structure ($\widehat{\mathbf{R}}$). (c) scaled noise structure ($\mathbf{E}$). For visualizations, subjects in scanner j were reordered by applying hierarchical clustering to $\widehat{\mathbf{I}}_j$, and imaging features were reordered by applying hierarchical clustering to $\widehat{\mathbf{I}}$.

Figure 3 contains the heatmaps of 3 components of UNIFAC harmonization. The heatmap of $\widehat{\mathbf{I}}$ in Figure 3 presents clearly distinct variations specific to Siemens Prisma. On the contrary, the heatmap of $\widehat{\mathbf{R}}$ does not show such pattern associated with scanner type, which also supports that the shared variations are irrelevant with scanner information. In addition, the heatmap of the rescaled noise does not show any noticeable patterns that could potentially affect homoscedasticity assumption.

To evaluate if scanner-specific latent patterns are well-removed, we computed the empirical covariances by scanners as well the difference between two scanner-specific covariances. The Figure 4 shows that the covariance differences remain notable in Adjusted Residuals, ComBat and CovBat harmonized data. SLIDE and AJIVE performed slightly better than ComBat and

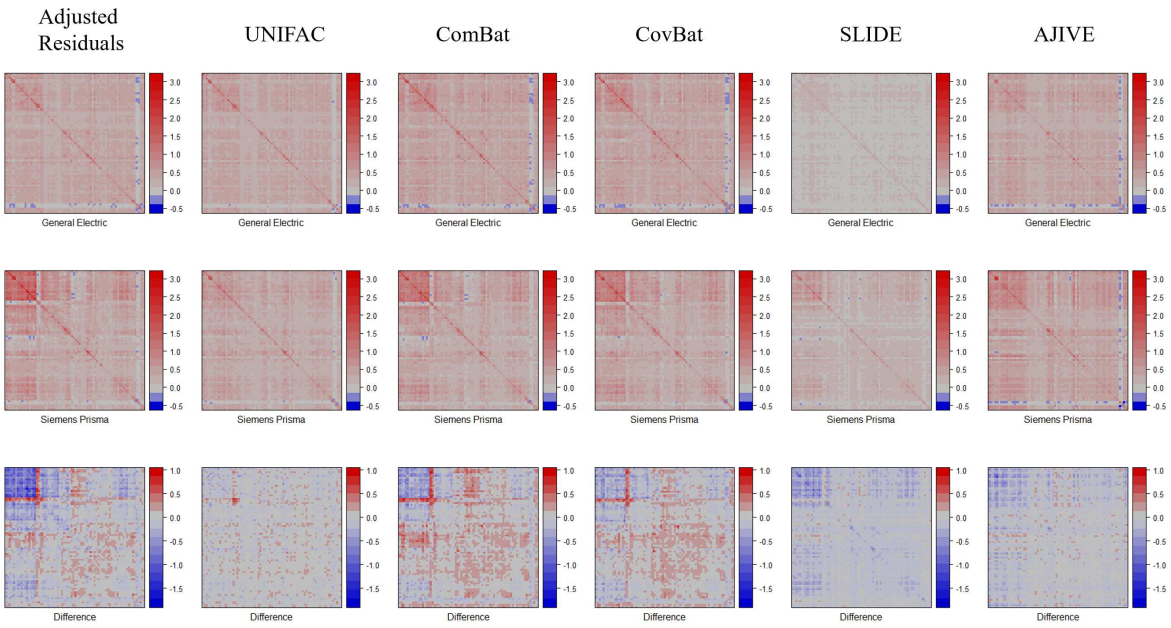


Fig. 4: Covariance matrices for harmonized data acquired from two scanners and their difference. All covariance matrices are estimated after residualizing the data on all potential covariates. The order of the feature agrees with Figure 3. The third row represents the difference of scanner-specific covariance matrices (first and second rows).

CovBat in mitigating covariance scanner effect. However, these covariance differences are considerably reduced with UNIFAC harmonization. We also assessed the quantitative comparisons for Frobenius norm of the scanner-specific covariance matrices. The norm for the UNIFAC harmonization was the lowest (5.09) followed by SLIDE (7.89), CovBat (8.73), AJIVE (8.84), and ComBat (9.61), which suggests superior performance of UNIFAC harmonization in constructing homogeneous covariances.

We also used Quadratic Discriminant Analysis (QDA) to evaluate how each harmonized data predicts scanners. A harmonization method that performs better in removing scanner effects will result in a lower prediction accuracy. Using machine learning methods to predict scanners from harmonized data has been used in previous literature (Fortin *and others*, 2018; Chen *and others*, 2022). Among existing classifiers, we chose QDA because the classifier is constructed

based on the mean vectors and covariance matrices only. Using leave-one-out cross-validation, we computed the average accuracy for each harmonized data after regressing out covariate effects. The UNIFAC harmonization method achieved the lowest prediction accuracy (49.6%) close to a random prediction, followed by CovBat (59.3%), ComBat (66.1%), SLIDE (66.4%) and AJIVE (68.4%).

Lastly, we investigated whether UNIFAC harmonization preserves the biological variability in the data. This step is necessary because the multivariate harmonization methods are prone to potentially overkilling too much variations including signals of interest. Here, we conducted parametric bootstrap to evaluate the power of the different harmonization methods. For the bth bootstrap, the procedure is summarized by (i) estimating all components of UNIFAC harmonization $(\widehat{\mathbf{R}}, \widehat{\mathbf{I}}, \hat{\delta}, \hat{\boldsymbol{\sigma}})$ using DTI-FA data, where $\hat{\boldsymbol{\sigma}}$ is the sample standard deviation vector for residuals from the regression with covariates, and computing sample covariance matrices of $\widehat{\mathbf{R}}, \widehat{\mathbf{I}}_1, \widehat{\mathbf{I}}_2$, denoted by $\widehat{\boldsymbol{\Sigma}}_{\mathbf{R}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{I}_1}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{I}_2}$; (ii) generating $\mathbf{I}_j^{(b)}$ from $\mathcal{MVN}(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{I}_j})$ $(j = 1, 2)$; $\mathbf{R}^{(b)}$ from $\mathcal{MVN}(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{R}})$; and each element of $\mathbf{E}^{(b)}$ from $\mathcal{N}(0, 1^2)$; (iii) generating $\mathbf{Y}^{(b)} = c \times \mathbf{I}^{(b)} + \mathbf{R}^{(b)} + \hat{\boldsymbol{\sigma}}\mathbf{1}^T \circ [\hat{\delta}_1 \mathbf{E}_1^{(b)}; \hat{\delta}_2 \mathbf{E}_2^{(b)}]$, where $\mathbf{1}$ is vector of ones and $\circ$ is the element-wise product. Here, we chose $c \in \{1, 2, 3\}$ to evaluate power based on different degrees of scanner-specific latent patterns. We then (iv) apply each harmonization method and build regression between each DTI-FA feature in each harmonized data and covariates $\mathbf{X}$, conduct t-test to the estimated coefficient of age×disease status, and extract the smallest p-value among all features. We repeat (ii)-(iv) B times to compute the power.

We evaluated power with respect to different Type 1 error threshold $\alpha \in (0, 0.1)$, which is shown in the Figure 5. UNIFAC harmonization gained more power than other methods, which indicates UNIFAC harmonization can retain more biological associations of interest than them so that it performed better in detecting true biological effects in the data. In addition, when we increased c, the relative weight of individual latent pattern ($\mathbf{I}$) in the bootstrap samples,

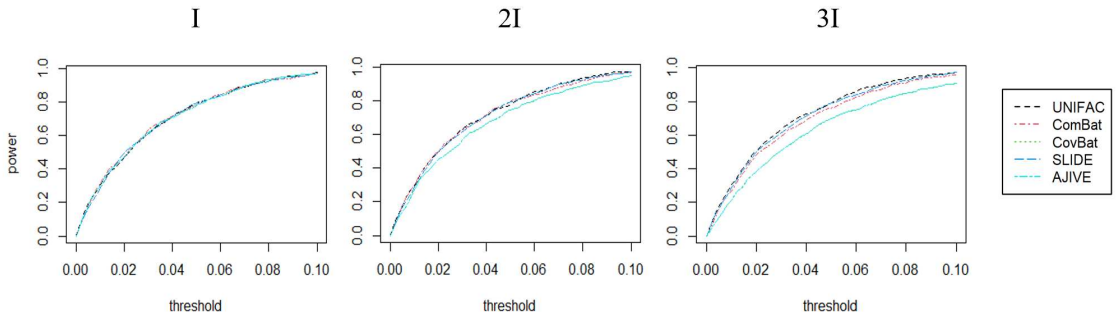


Fig. 5: Summary of power for five harmonization methods. The plots from left to right are with increased proportion of individual latent pattern

the performance of UNIFAC harmonization remained similar but the power of other methods decreased. The oversimplified model assumption for ComBat, that scanner effects only account for location and scales, could result in scanner effects still remaining in the resulting data. These effects could overshadow our detection of biological associations. On the contrary, the CovBat's performances indicate a loss of power when the heterogeneous covariance assumption set by CovBat is not met.

4. Simulation Studies

4.1 *Simulation designs*

We performed extensive simulation studies to evaluate the performance of UNIFAC harmonization and to compare it to other methods. We included ComBat, CovBat, SLIDE, AJIVE and Adjusted Residuals as our competitors and then evaluated whether harmonized data preserves biological variations through the power analysis of harmonized data. To evaluate the control of false positives and power, we designed two simulation experiments. We considered $J = 2$ in this framework, because it is consistent with our data analysis in Section 3 and allows to consider SLIDE method in the evaluation framework.

Simulation 1: We generated data using the sum of low-rank features. We simulated 1,000 null data sets with $n_1 = n_2 = 50$ (so that $n = 100$), and $p = 100$ features. Our data generating model is summarized by:

$$\mathbf{Y} = \underbrace{\boldsymbol{\beta}\mathbf{X}'}_{\text{rank } 4} + \underbrace{\boldsymbol{\Gamma}}_{\text{rank}2} + \underbrace{\mathbf{R}}_{\text{rank } 3} + \underbrace{c \cdot \mathbf{I}}_{\text{rank } 6} + [\delta_1 \mathbf{E}_1, \ldots, \delta_J \mathbf{E}_J].$$

We first used 4 nuisance covariates for covariate effects, where each element of $\boldsymbol{\beta}$ was generated from $\mathcal{N}(0, 1^2)$. The covariate vector for each subject was generated from the multivariate normal distribution with zero mean, and we used AR1(0.2) for the covariance matrix. Second, we generated $\mathbf{R}$ by first generating a $p \times n$ matrix whose entries are drawn from $\mathcal{N}(0, 1^2)$, then taking the first 3 principal components. Similarly, we generated each $\mathbf{I}_j$ by generating a $p \times n_j$ matrix using $\mathcal{N}(0, 1^2)$ then taking the top 3 principal components. Lastly, we also generated the additive batch effect (location) γ_{ij} by fixing it be the same for all i and from $\mathcal{N}(0, 1.5^2)$, and multiplicative batch effect (scale) δ_j from Uniform$(1, 1.5)$. Finally, the elements of $\mathbf{E}$ were generated from $\mathcal{N}(0, 1^2)$.

We note that, c was chosen between $0, 1, 2, 3$ to evaluate the impact of scanner-specific latent patterns on statistical power. Note that we also considered $c = 0$ to investigate whether it has comparable performance when data generating model does not include any batch-specific latent patterns.

Simulation 2: We generated data by modifying the simulation design introduced by Chen *and others* (2022). To address potential covariance batch effects, the CovBat model uses principal component (PC) scores to shift each within-batch covariance to the pooled covariance structure. Therefore, the design aimed at evaluating whether harmonization methods can approximate the underlying covariance structure when covariance batch effects are captured by its PC shifts.

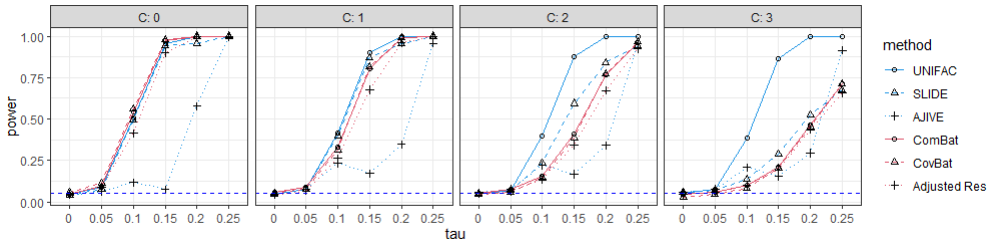
We simulated 1,000 null data sets based on SPINS data so that $n_1 = 172, n_2 = 179$ (so that $n = 351$) and $p = 73$ features. The data y_{ijk} was generated by $y_{ijk} = \alpha_i + \gamma_{ij} + \delta_{ij}\epsilon_{ijk}$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$ is the sample mean vector of Scanner General Electric observations in the SPINS data. The additive scanner effects $\boldsymbol{\gamma}_j = (\gamma_{1j}, \ldots, \gamma_{pj})^T$'s are vectors drawn from

$\mathcal{N}(0, 0.1^2)$. For multiplicative scanner effects, we used $\delta_{i1} \sim \mathcal{IG}(46, 50)$ and $\delta_{i2} \sim \mathcal{IG}(51, 50)$. From the sample correlation matrix of DTI-FA observations in the SPINS data $\mathbf{S}$ with the corresponding eigen decomposition $\mathbf{S} = \sum_{l=1}^{73} \hat{\lambda}_l \hat{\boldsymbol{\psi}}_l \hat{\boldsymbol{\psi}}_l^T$, we generated $\boldsymbol{\epsilon}_{jk} = (\epsilon_{1jk}, \ldots, \epsilon_{pjk})^T$ that contained scanner-specific shifts in two designs. The first design was introduced by Chen *and others* (2022) to investigate how the rank of the covariance effect influences harmonization results, and we generated error terms by $\boldsymbol{\epsilon}_{jk} \sim \mathcal{MVN}(\mathbf{0}, \mathbf{S} + c_j \sum_{l=1}^{K} \hat{\lambda}_l \hat{\boldsymbol{\psi}}_l \hat{\boldsymbol{\psi}}_l^T)$, where $c_1 = -\frac{1}{2}$ and $c_2 = \frac{1}{2}$. We considered different K including $K = 0, 1, 5, 10$. The second design aimed to investigate how the severity of the covariance shift influences harmonization results, and we generated $\boldsymbol{\epsilon}_{jk} \sim \mathcal{MVN}(\mathbf{0}, c_j \sum_{l=1}^{K} \hat{\lambda}_l \hat{\boldsymbol{\psi}}_l \hat{\boldsymbol{\psi}}_l^T + \sum_{l=K+1}^{73} \hat{\lambda}_l \hat{\boldsymbol{\psi}}_l \hat{\boldsymbol{\psi}}_l^T)$, where $K = 5$, $c_1 = 0.125, 0.25, 0.5, 0.75$ and $c_2 = 1.875, 1.75, 1.5, 1.25$.

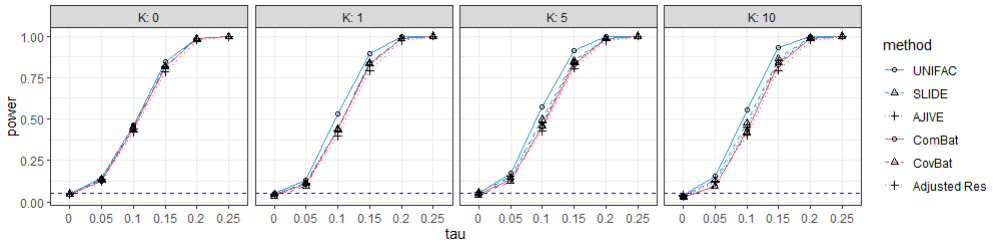
For evaluation of power, we generated our covariate of interest, Z_k $(k = 1, \ldots, n)$, randomly from 0 or 1. We then randomly chose 20% (from Simulation 1) and 50% (from Simulation 2) of features and added $\tau_i \cdot Z_k$ to the null data, where $\tau_i > 0$ is the effect size for the ith feature. For both simulation setups, we used permutation to control family-wise error rate (FWER) and evaluated the power for the null hypothesis $H_0 : \tau_i = 0$ for all $i = 1, \ldots, p$, controlled at the level of 5%.

4.2 *Simulation results*

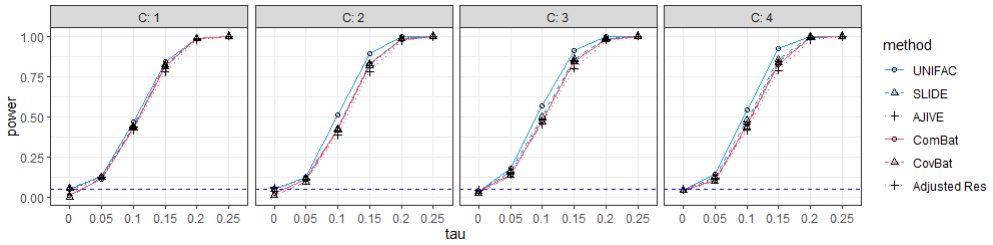
The results for Simulation 1 are summarized in Figure 6(i). UNIFAC harmonization controlled family-wise error properly, with empirical FWER of 0.044, 0.047, 0.048, and 0.052 regardless of the choice of c. In our simulations, while other methods controlled FWER appropriately in most scenarios, CovBat was conservative in controlling false positives when the proportion of individual latent patterns increased. In terms of power, UNIFAC harmonization's performance was nearly the same as ComBat or CovBat even when the data generating model is misspecified (i.e., $c = 0$), which shows the robustness of the proposed method. Furthermore, as c increased,



(i) The plots from left to right are with increased proportion of individual latent pattern.



(ii) The plots from left to right are with increased rank of the covariance effect.



(iii) The plots from left to right are with decreased severity of the covariance shift.

Fig. 6: Summary of power for six harmonization methods. The blue dashed horizontal line is FWER=0.05

UNIFAC harmonization showed substantial power gain compared to others, partially because it correctly identified and removed the batch-specific latent patterns in the data. The lower power of ComBat and Adjusted Residuals are expected as they do not consider these latent patterns in their model, and the lower power of CovBat is also expected because the data generating model is different from the CovBat's assumption on PC shifts. Although AJIVE and SLIDE take these latent patterns into consideration, their power was lower than UNIFAC harmonization when c is large. It might be because UNIFAC's objective function was superior in signal reconstructions with various degrees of signal-to-noise ratios (Park and Lock, 2020).

The results for Simulation 2 are summarized in Figure 6(ii) and 6(iii). In Design 1, UNIFAC harmonization's empirical FWER are 0.04, 0.044, 0.06, 0.053 for $K = 0, 1, 5, 10$, while ComBat, CovBat and SLIDE are conservative in controlling false positives when K was small. For power, we note that when the null hypothesis was not true, all harmonization methods increased statistical power and performed similarly when K is small. When K was large, UNIFAC harmonization still showed superior performance to other methods which supports the efficiency of UNIFAC harmonization even when the data generating model did not follow the assumption of UNIFAC harmonization. In Design 2, UNIFAC harmonization controlled family-wise error properly, with empirical FWER of 0.05, 0.052, 0.041, 0.047. On the contrary, empirical FWER of ComBat were 0.018, 0.029, 0.037, 0.044, and of CovBat were 0.003, 0.016, 0.028, 0.045, making them very conservative in controlling false positives when the severity of covariance shift was high (i.e., $c_1 = 0.125, c_2 = 1.875$). When the severity of covariance shift was high, all harmonization methods showed similar performance in increasing statistical power in the hypothesis test. However, when the severity of covariance shift was low/moderate, UNIFAC harmonization showed superior performance over other methods.

5. Discussion

We proposed a novel harmonization method, called UNIFAC harmonization, that estimates and removes both explicit (additive and multiplicative) and latent batch effects. We also provided a framework to reformulate existing joint and individual factorization methods in the data harmonization context and compared their performances. While multivariate harmonization itself has been proposed in the literature, our approach that models batch-specific latent patterns provides an intuitive and interpretable way to view the problem as heterogeneous covariances when the low-rank assumption is satisfied. This novel method provides a new direction in neuroimaging and genomics that satisfies the growing need of correcting batch effects for multi-site/scanner

studies.

We showed in analysis of SPINS data and simulations that UNIFAC harmonization is superior to other methods in harmonizing *covariances* across batch effects while retaining (biological) variations unnrelated to batches and improving reproducibility. In SPINS data analysis, we compared these methods for harmonizing mean fractional anisotropy (FA) measurements of DTI images. Our data analysis showed that UNIFAC harmonization performed better in entirely removing batch effects as well as retaining associations of interest to increase statistical power than other methods. In simulation studies, we showed UNIFAC harmonization maintained FWER control for multiple comparisons properly and had superior performance in increasing statistical power compared to other methods. The difference in the statistical power between UNIFAC harmonization and other methods was notable as the proportion of latent batch effects increased.

We describe some limitations of UNIFAC harmonization. Our current approach is evaluated with a moderate number of features and samples, and its performance in high dimensional features ($p \gg n$) has not been explored. Also, UNIFAC harmonization assumes that the original data matrix consists of low rank signals (including latent batch effects) plus full rank noises to scale data and choose tuning parameters. Its performance can be affected if this assumption does not hold. For example, vertex-wise cortical thickness data has at most 160,000 features in each hemishpere of the brain and reveals a high degree of spatial autocorrelation that low-rank assumption is not reasonable. Also, we currently use random matrix theory to select tuning parameter. Even if low-rank assumption holds, empirical approaches (e.g., cross-validation) may also be used to select tuning parameters at the expense of computational cost, which may potentially improve the quality of harmonized data (Owen and Perry, 2009). In addition, confounding between batches and biological covariates poses additional challenges for all harmonization methods, since removing unwanted variations associated batches may overkill biological associations. (Fortin *and others*, 2017; Nygaard *and others*, 2016). It may be a problem for UNIFAC harmonization as well, and

we leave the evaluation for it as future work. Lastly, we assumed normality of the data, which is well-justified in neuroimaging data or log-transformed microarray data, but other data types in genomics (e.g. RNA-Seq) that reveal excessive zero counts will require a different probabilistic model for latent batch effects. We leave it as future work.

The UNIFAC haromonization has room for improvement. First, UNIFAC harmonization is applied to general multivariate data, and it might also be extended to accommodate different data types' needs, i.e., structural, functional, and other imaging modalities. It would be interesting to explore if latent batch effects are shared across data modalities. Also, it would be interesting to explore whether the non-biological variations we detected (i.e., $\mathbf{I}$) can be explained by existing non-biological information (e.g, scanner information), which, if it exists, would be helpful in validation of the harmonization method.

The proposed method is publicly available as a R package at `https://github.com/junjypark/UNIFACharmonization`.

Acknowledgements

The authors would like to thank all participants for their contribution to this work, and the research staff who performed data collection and management. The SPINS study was supported by the National Institute of Mental Health (1/3R01MH102324-01, 2/3R01MH102313-01, 3/3R01MH102318-01). LDO was supported by the Brain & Behavior Research Foundation. ANV was supported by the National Institute of Mental Health (1/3R01MH102324 & 1/5R01MH114970), Canadian Institutes of Health Research, Canada Foundation for Innovation, CAMH Foundation, and University of Toronto. JYP was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN-2022-04831) and the University of Toronto's Data Science institute (Catalyst Grant).

References

Beaulieu, Christian. (2002). The basis of anisotropic water diffusion in the nervous system–a technical review. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo* **15**(7-8), 435–455.

Chen, Andrew A, Beer, Joanne C, Tustison, Nicholas J, Cook, Philip A, Shinohara, Russell T, Shou, Haochang and Initiative, Alzheimer's Disease Neuroimaging. (2022). Mitigating site effects in covariance for machine learning in neuroimaging data. *Human brain mapping* **43**(4), 1179–1195.

Feng, Qing, Jiang, Meilei, Hannig, Jan and Marron, JS. (2018). Angle-based joint and individual variation explained. *Journal of multivariate analysis* **166**, 241–265.

Fortin, Jean-Philippe, Cullen, Nicholas, Sheline, Yvette I, Taylor, Warren D, Aselcioglu, Irem, Cook, Philip A, Adams, Phil, Cooper, Crystal, Fava, Maurizio, McGrath, Patrick J *and others*. (2018). Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* **167**, 104–120.

Fortin, Jean-Philippe, Parker, Drew, Tunç, Birkan, Watanabe, Takanori, Elliott, Mark A, Ruparel, Kosha, Roalf, David R, Satterthwaite, Theodore D, Gur, Ruben C, Gur, Raquel E *and others*. (2017). Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* **161**, 149–170.

Fortin, Jean-Philippe, Sweeney, Elizabeth M, Muschelli, John, Crainiceanu, Ciprian M, Shinohara, Russell T, Initiative, Alzheimer's Disease Neuroimaging *and others*. (2016). Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage* **132**, 198–212.

Gagnon-Bartsch, Johann A, Jacob, Laurent and Speed, Terence P. (2013). Removing

unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California*, 1–112.

Gavish, Matan and Donoho, David L. (2017). Optimal shrinkage of singular values. *IEEE Transactions on Information Theory* **63**(4), 2137–2152.

Gaynanova, Irina and Li, Gen. (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics* **75**(4), 1121–1132.

Johnson, W Evan, Li, Cheng and Rabinovic, Ariel. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**(1), 118–127.

Leek, Jeffrey T and Storey, John D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* **3**(9), e161.

Lock, Eric F, Hoadley, Katherine A, Marron, James Stephen and Nobel, Andrew B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The annals of applied statistics* **7**(1), 523.

Lock, Eric F, Park, Jun Young and Hoadley, Katherine A. (2022). Bidimensional linked matrix factorization for pan-omics pan-cancer analysis. *The annals of applied statistics* **16**(1), 193.

Malcolm, James G, Shenton, Martha E and Rathi, Yogesh. (2010). Filtered multitensor tractography. *IEEE transactions on medical imaging* **29**(9), 1664–1675.

Nygaard, Vegard, Rødland, Einar Andreas and Hovig, Eivind. (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17**(1), 29–39.

Owen, Art B and Perry, Patrick O. (2009). Bi-cross-validation of the SVD and the non-negative matrix factorization. *The annals of applied statistics* **3**(2), 564–594.

Park, Jun Young and Lock, Eric F. (2020). Integrative factorization of bidimensionally linked matrices. *Biometrics* **76**(1), 61–74.

Yu, Meichen, Linn, Kristin A, Cook, Philip A, Phillips, Mary L, McInnis, Melvin, Fava, Maurizio, Trivedi, Madhukar H, Weissman, Myrna M, Shinohara, Russell T and Sheline, Yvette I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human brain mapping* **39**(11), 4213–4227.

Zhang, Yuqing, Parmigiani, Giovanni and Johnson, W Evan. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR genomics and bioinformatics* **2**(3), lqaa078.