Integrative Factor Regression and Its Inference for Multimodal Data Analysis

Quefeng Li and Lexin Li

University of North Carolina, Chapel Hill and University of California, Berkeley

Abstract

Multimodal data, where different types of data are collected from the same subjects, are fast emerging in a large variety of scientific applications. Factor analysis is commonly employed in integrative analysis of multimodal data, and is particularly useful to overcome the curse of high dimensionality and high correlations of multimodal data. However, there is little work on statistical inference for factor analysis based supervised modeling of multimodal data. In this article, we consider an integrative linear regression model that is built upon the latent factors extracted from multimodal data. We address three important questions: how to infer the significance of one data modality given the other modalities in the model; how to infer the significance of a combination of variables from one modality or across different modalities; and how to quantify the contribution, measured by the goodness-of-fit, of one data modality given the others. When answering each question, we explicitly characterize both the benefit and the extra cost of factor analysis. Those questions, to our knowledge, have not yet been addressed despite wide use of factor analysis in integrative multimodal analysis, and our proposal thus bridges an important gap. We study the empirical performance of our methods through simulations, and further illustrate with a multimodal neuroimaging analysis.

Keywords: Data integration; Dimension reduction; Factor analysis; High-dimensional inference; Multimodal neuroimaging; Principal components analysis.

1 Introduction

Thanks to rapid technological advances, multiple types of data are now frequently collected for a common set of experimental subjects. Such a new data structure, often referred as multi-view, multi-source or multimodal data, is fast emerging in a wide range of scientific fields. Examples include multi-omics data in genomics, multimodal neuroimaging data in neuroscience, multimodal electronic health records data in health care administration, among others. Numerous empirical studies have found that, by combining diverse but usually complementary information from different types of data, an integrative analysis of multimodal data is often beneficial; see Uludag and Roebroeck (2014); Li et al. (2016); Richardson et al. (2016) for reviews and the references therein.

In view of the promise of multimodal data, a number of statistical methods have recently been developed for integrative analysis. An important class of such solutions is matrix or tensor factorization, which decomposes multimodal data into the components that capture joint variation shared across modalities, and the components that characterize modality-specific variation (Lock et al., 2013; Yang and Michailidis, 2015; Li and Jung, 2017; Lock and Li, 2018; Gaynanova and Li, 2019). Another class is canonical correlation analysis, which seeks maximum correlations between different data modalities through decomposition of the between-modality dependency structure (Li and Gaynanova, 2018; Shu et al., 2019). However, all these methods are unsupervised, in the sense that there is not a response variable involved. Li et al. (2018) recently proposed an integrative reduced-rank regression to model multivariate responses given multi-view data as predictors. Xue and Qu (2019) developed an estimating equations approach to accommodate block missing patterns in multimodal data. Their methods are supervised, but both focused on parameter estimation and variable selection instead of statistical inference.

It is of ubiquitous interest to study the predictive associations between responses and multimodal predictors. However, there are some unique characteristics of multimodal data that make the problem challenging. First, multimodal data are often high-dimensional. In plenty of applications, even a single modality contains more variables than the sample size. Second, multimodal data are often highly correlated, as they measure related features of the same subjects and thus often share common variations across different modalities.

This phenomenon has been constantly observed, and is actually the base upon which those matrix or tensor factorization solutions are built (Lock et al., 2013). Such high correlations render many standard high-dimensional models, e.g., LASSO (Tibshirani, 1996), unsuitable, as they usually require the predictors not to be highly correlated in order to achieve the desired statistical properties. Finally, multimodal data pose new questions; for instance, how to quantify the contribution and statistical significance of one data modality conditioning on the other data modalities in the regression model.

Factor analysis is a well-known approach to both reduce high dimensionality and high correlations among the variables. For data with a single modality, Fan et al. (2013) employed a factor model to estimate a non-sparse covariance matrix. Kneip et al. (2011) proposed to include the latent factors as additional explanatory variables in a high-dimensional linear regression, and established the model selection consistency. Fan et al. (2016) proposed a factor-adjusted model selection method for a general high-dimensional M-estimation problem. They separated the latent factors from the idiosyncratic components to reduce correlations among the covariates, and showed that their method can reach the variable selection consistency under milder conditions than standard selection methods. Li et al. (2018) studied estimation of a covariance matrix of variables. They showed that leveraging on additional auxiliary variables can improve the estimation, when the auxiliary variables share some common latent factors with the variables of interest. For data with multiple modalities, Shen et al. (2013) proposed an integrative clustering method based on identifying common latent factors from multi-omics data. Zhang et al. (2019) developed an imputed factor regression model for dimension reduction and prediction of multimodal data with missing blocks. Despite these efforts, however, there is little work on statistical inference for supervised modeling of multimodal data. Moreover, there is no explicit quantification of the benefit of factor analysis in a multimodal regression setting, and many important inference-related questions remain unanswered.

In this article, we aim to bridge this gap. We consider an integrative linear regression model built upon the latent factors extracted from multimodal data. We show that this model alleviates high dimensionality and high correlations of multimodal data. Based on this model, we address three important questions: how to infer the significance of one data modality given the other modalities in the model; how to infer the significance of a combination of variables from one or more modalities; and how to quantify the contribution, measured by the goodness-of-fit, of one data modality given the others. When answering each question, we explicitly characterize both the benefit and the extra cost of factor analysis. First, by resorting to a relatively small number of latent factors, it effectively reduces the dimensionality and turns a high-dimensional test to a low-dimensional one when testing the significance of a whole modality. As a result, it enables us to derive a closed form of the limiting distribution of the test statistic; see Theorem 1. Second, our method can consistently estimate the support and nonzero components of the covariate coefficients in the regression model. More importantly, by using the decorrelated idiosyncratic components from factor analysis as the pseudo predictors, instead of the original highly correlated covariates, it requires much weaker conditions to reach the variable selection and estimation consistency; see Theorem 2. Moreover, we show that, when there are enough variables in each modality so that the latent factors can be well estimated, the resulting estimation error can reach the minimax optimal rate, but again under weaker conditions. Third, such an improvement in selection and estimation in turn benefits the inference of the significance of a linear combination of predictors, by requiring less stringent conditions to establish the limiting distribution of the test statistic; see Theorem 3. Finally, by leveraging on the latent factors shared across modalities, it enables us to obtain a closed-form measure of the variance of the response explained by one modality in addition to the others. Such a measure facilitates the quantification of the contribution of an individual modality.

Our proposal contributes on several fronts. Even though factor analysis has been widely used in multimodal data analysis, there has been no formal test developed to explicitly quantify the contribution and significance of an individual modality or a related set of variables across different modalities. Our proposal provides the first inferential tools to address those important questions. Moreover, our work is built on careful examination of the benefit and trade-off of factor analysis in regression. Compared to the existing literature, the proof techniques are much more involved than those of the standard setting when the design matrix is observed and fixed with a single data modality. The technical tools we develop here are not limited to our setting alone, but are applicable to general supervised high-dimensional factor models. We also remark that, although we focus on a linear factor regression model, most of the inference-related results we obtain can be extended to a general M-estimation problem such as a generalized linear model.

We employ the following notation throughout this article. For a vector $\mathbf{a} \in \mathcal{R}^d$, let $\|\mathbf{a}\|_\infty = \max_j |a_j|$, $\|\mathbf{a}\|_1 = \sum_{j=1}^d |a_j|$, $\|\mathbf{a}\|_2 = (\sum_{j=1}^d a_j^2)^{1/2}$ denote its sup-norm, L_1-norm and Euclidean norm, respectively. For an index set S, let $\mathbf{a}_S$ denote the subvector of $\mathbf{a}$ with indices in S. In particular, let the subscript m denote the index set of the mth modality, and the subscript $-m$ denote the index set of all other modalities. Let $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}'$ denote its outer product. Let $\text{supp}(\mathbf{a}) = \{j : a_j \neq 0\}$ denote the support of $\mathbf{a}$. For a square matrix $\mathbf{A} = (a_{ij}) \in \mathcal{R}^{d\times d}$, let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote its minimum and maximum eigenvalues. Let $\|\mathbf{A}\|_\infty = \sup_{ij} |a_{ij}|$, $\|\mathbf{A}\|_1 = \max_{1\le j\le d} \sum_{i=1}^d |a_{ij}|$, $\|\mathbf{A}\|_2 = \lambda_{\max}(\mathbf{A})$, $\|\mathbf{A}\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$ denote its element-wise sup-norm, L_1-norm, L_2-norm, and Frobenious norm, respectively. Let $\mathbf{A}_S$ denote the submatrix of $\mathbf{A}$ with row and column indices in S. For a rectangular matrix $\mathbf{B} = (b_{ij}) \in \mathcal{R}^{m\times n}$, let $\|\mathbf{B}\|_{L_\infty} = \max_{1\le i\le m} \sum_{j=1}^n |b_{ij}|$. For two sequences a_n and b_n, write $a_n = o(b_n)$ if $a_n/b_n \to 0$, and $a_n \gg b_n$ if $b_n/a_n \to 0$. For an integer M, let $[M] = \{1, \ldots, M\}$. For a set S, let $|S|$ denote the number of elements in S.

The rest of the article is organized as follows. We introduce the integrative factor regression model in Section 2, and describe the parameter estimation in Section 3. These results are mostly built upon the existing literature on factor analysis. Then we address the three questions, which to our knowledge have not been answered before. That is, we develop a test to evaluate the significance of an individual modality given the other modalities in Section 4, develop a test for a linear combination of predictors in Section 5, and derive a measure to quantify the contribution of an individual modality in Section 6. We present the simulations and a multimodal neuroimaging data example in Section 7.

2 Integrative factor regression model

Suppose there are M modalities of variables. Let $\mathbf{x}_m \in \mathcal{R}^{p_m}$ denote the vector of p_m random variables from the mth modality, and y denote the response variable. Let $\mathbf{x} = (\mathbf{x}_1', \ldots, \mathbf{x}_M')' \in \mathcal{R}^p$, and $p = \sum_{m=1}^M p_m$. We assume $\mathbf{x}_m$ is driven by some latent factors in that $\mathbf{x}_m$ can be decomposed as

$$\mathbf{x}_m = \mathbf{\Lambda}_m \mathbf{f}_m + \mathbf{u}_m, \tag{1}$$

where $\mathbf{f}_m \in \mathcal{R}^{K_m}$ is the vector of K_m random latent factors, $\mathbf{u}_m \in \mathcal{R}^{p_m}$ is the vector of random idiosyncratic errors of variables in the mth modality that are uncorrelated with $\mathbf{f}_m$, and $\mathbf{\Lambda}_m \in \mathcal{R}^{p_m \times K_m}$ is the loading matrix of $\mathbf{x}_m$ on the latent factors $\mathbf{f}_m$. To avoid the identifiability issue on $\mathbf{\Lambda}_m$ and $\mathbf{f}_m$, we adopt the usual assumption in the factor analysis literature by assuming that

$$\mathrm{Var}(\mathbf{f}_m) = \mathbf{I}_{K_m}, \quad \text{and} \quad \mathbf{\Lambda}_m' \mathbf{\Lambda}_m = \mathbf{D}_m \in \mathcal{R}^{K_m \times K_m} \text{ is a diagonal matrix.}$$

We also assume that $\mathbf{f} = (\mathbf{f}_1', \ldots, \mathbf{f}_M')' \in \mathcal{R}^K$ is uncorrelated with $\mathbf{u} = (\mathbf{u}_1', \ldots, \mathbf{u}_M')' \in \mathcal{R}^p$, where $K = \sum_{m=1}^M K_m$. Let $\mathbf{\Lambda} = \mathrm{diag}(\mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_M) \in \mathcal{R}^{p \times K}$ be the block diagonal matrix of the loading matrices from all modalities, and $\mathbf{\Sigma}_u = \mathrm{diag}(\mathbf{\Sigma}_{u_1}, \ldots, \mathbf{\Sigma}_{u_M})$ be the block diagonal matrix of the covariance of the idiosyncratic errors. In the factor analysis literature, it is often assumed that $\mathbf{\Sigma}_u$ is sparse, which means that, after removing the variations contributed by the latent factors, the correlations among the idiosyncratic components are weak. Therefore, the idiosyncratic $\mathbf{u}$ can be viewed as a decorrelated version of the original variables $\mathbf{x}$.

We next employ a linear model to connect $\mathbf{x}_m$ with y, in that,

$$y = \sum_{m=1}^{M} \mathbf{x}_m' \boldsymbol{\beta}_m^* + \epsilon, \tag{2}$$

where $\boldsymbol{\beta}_m^* \in \mathcal{R}^{p_m}$ is the true effect of $\mathbf{x}_m$ on the response y, and ϵ is an error independent of $\mathbf{x}_m$ with $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma_\epsilon^2$.

Suppose we have n i.i.d. realizations of the data, $\mathbf{Y} = (y_1, \ldots, y_n)' \in \mathcal{R}^n$, $\mathbf{X}_m = (\mathbf{x}_{1,m}, \ldots, \mathbf{x}_{n,m})' \in \mathcal{R}^{n \times p_m}$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)' \in \mathcal{R}^n$. Then model (2) can be written as

$$\mathbf{Y} = \sum_{m=1}^{M} \mathbf{X}_m \boldsymbol{\beta}_m^* + \boldsymbol{\epsilon}.$$

By the factor model (1), we have $\mathbf{X}_m = \mathbf{F}_m \mathbf{\Lambda}_m' + \mathbf{U}_m$, where $\mathbf{F}_m = (\mathbf{f}_{1,m}, \ldots, \mathbf{f}_{n,m})' \in \mathcal{R}^{n \times K_m}$ is the matrix of the K_m factors in the mth modality pertaining to the n subjects, and $\mathbf{U}_m = (\mathbf{u}_{1,m}, \ldots, \mathbf{u}_{n,m})' \in \mathcal{R}^{n \times p_m}$ is the matrix of idiosyncratic errors. Then, we have,

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\gamma}^* + \mathbf{U}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{3}$$

where $\mathbf{F} = (\mathbf{F}_1, \ldots, \mathbf{F}_M) \in \mathcal{R}^{n \times K}$, $\boldsymbol{\gamma}^* = (\boldsymbol{\beta}_1^{*\prime}\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\beta}_M^{*}{}'\boldsymbol{\Lambda}_M)' \in \mathcal{R}^K$, $\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_M) \in \mathcal{R}^{n \times p}$, and $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*\prime}, \ldots, \boldsymbol{\beta}_M^{*}{}')' \in \mathcal{R}^p$. We call model (3) an integrative factor regression model. In the remainder of this article, we aim to show that model (3) can benefit estimation, selection and inference about $\boldsymbol{\beta}^*$. The intuition is that, after the latent factors $\mathbf{F}$ are separated, the idiosyncratic error $\mathbf{U}$ can be treated as the pseudo predictors, whose between-variable correlation is much weaker than the correlation in the original $\mathbf{X}$. Such a decorrelation eases selection of $\boldsymbol{\beta}^*$, which in turn benefits the inference on $\boldsymbol{\beta}^*$. In addition, the factor decomposition also serves as a dimension reduction tool, which enables us to derive some closed-form results in inference.

3 Estimation

For the integrative factor regression model (3), our primary interest is to estimate and infer about $\boldsymbol{\beta}_m^*$, which reflects the effect of the mth modality $\mathbf{x}_m$ on the response y. To achieve that, we first estimate the latent variables $\mathbf{F}$ and $\mathbf{U}$, along with the number of latent factors K_m, using some well established methods in the factor analysis literature. We then estimate $\boldsymbol{\beta}_m^*$ through a penalized least squares approach.

First we estimate the latent variables $\mathbf{F}$ and $\mathbf{U}$. We adopt the method in Bai and Li (2012) and Fan et al. (2013), by running principal components analysis (PCA) on each individual modality $\mathbf{X}_m$. Other PCA methods such as Ma (2013); Cai et al. (2013) are also applicable here. We then estimate $\mathbf{F}_m$ by $\sqrt{n}$ times eigenvectors corresponding to the largest K_m eigenvalues of $\mathbf{X}_m\mathbf{X}_m'$. Denote this estimator by $\widehat{\mathbf{F}}_m$. We next estimate $\boldsymbol{\Lambda}_m$ by $\widehat{\boldsymbol{\Lambda}}_m = (1/n)\mathbf{X}_m'\widehat{\mathbf{F}}_m$, and estimate $\mathbf{U}_m$ by $\widehat{\mathbf{U}}_m = \mathbf{X}_m - \widehat{\mathbf{F}}_m\widehat{\boldsymbol{\Lambda}}_m'$, accordingly. Fan et al. (2013) showed that $\widehat{\mathbf{F}}_m$ is a consistent estimator, up to a rotation, of $\mathbf{F}_m$, under a pervasive condition that the latent factors should affect many variables; see Condition 2 and its discussion in Section 4.

Next we determine the number of the latent factors K_m in each modality, which is usually unknown in practice. We use the method of Bai and Ng (2002) to estimate K_m by

$$\widehat{K}_m = \underset{0 \le k \le \widetilde{M}}{\operatorname{argmin}} \ \log\left\{\frac{1}{np_m}\|\mathbf{X}_m - n^{-1}\widehat{\mathbf{F}}_{m_k}\widehat{\mathbf{F}}_{m_k}'\mathbf{X}_m\|_F^2\right\} + kg(n, p_m),$$

where $\widetilde{M}$ is a pre-defined upper bound on K_m, $\widehat{\mathbf{F}}_{m_k}$ is $\sqrt{n}$ times eigenvectors corresponding to the largest m_k eigenvalues of $\mathbf{X}_m\mathbf{X}_m'$, and $g(n, p_m)$ is a penalty function that,

$$g(n, p_m) = \frac{n + p_m}{np_m} \log\left(\frac{np_m}{n + p_m}\right), \quad \text{or} \quad g(n, p_m) = \frac{n + p_m}{np_m} \log\left(\min\{n, p_m\}\right).$$

For both choices, Bai and Ng (2002) showed that $\widehat{K}_m$ is a consistent estimator of K_m under some regularity conditions. We make two additional remarks about K_m. First, in this article, we treat K_m as fixed, which is reasonable in numerous scientific applications. As K_m is related to the number of spiked eigenvalues of $\mathbf{X}_m\mathbf{X}_m'$, it is usually small. Second, in our subsequent theoretical analysis, we treat K_m as known. But all the theoretical results remain valid as long as K_m is replaced by a consistent estimator $\widehat{K}_m$.

Next we estimate $\boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}^*$. We replace $\mathbf{F}$ and $\mathbf{U}$ with the corresponding estimators $\widehat{\mathbf{F}} = (\widehat{\mathbf{F}}_1, \ldots, \widehat{\mathbf{F}}_M)$ and $\widehat{\mathbf{U}} = (\widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_M)$, and solve a penalized least squares problem,

$$(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}) = \underset{(\boldsymbol{\gamma},\boldsymbol{\beta})}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \widehat{\mathbf{F}}_i'\boldsymbol{\gamma} - \widehat{\mathbf{U}}_i'\boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} p(|\beta_j|), \tag{4}$$

where $p(\cdot)$ is some general folded-concave penalty function, and λ is a tuning parameter. This class of penalty functions includes SCAD (Fan and Lv, 2011) and MCP (Zhang, 2010). It assumes that $p(t)$ is increasing and concave in $t \geq 0$, and has a continuous first derivative $\dot{p}(t)$ with $\dot{p}(0+) > 0$. This optimization problem can be solved by standard proximal gradient descent algorithms (Parikh et al., 2014). We later show in Theorem 2 that the estimator $\widehat{\boldsymbol{\beta}}$ from (4) can consistently select the support and estimate the nonzero components of $\boldsymbol{\beta}^*$. We tune λ using the standard cross-validation method.

Finally we estimate σ_ϵ^2 by $\widehat{\sigma}_\epsilon^2 = (1/n)\sum_{i=1}^n (y_i - \widehat{\mathbf{F}}_i'\widehat{\boldsymbol{\gamma}} - \widehat{\mathbf{U}}_i'\widehat{\boldsymbol{\beta}})^2$, where $(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}})$ are obtained from (4). Alternatively, we can estimate σ_ϵ^2 by scaled LASSO (Sun and Zhang, 2013) or refitted cross-validation (Fan et al., 2012). All these are consistent estimators of σ_ϵ^2, which ensure the results in Theorems 2 and 3 hold.

4 Hypothesis test of a whole modality

A crucial question in multimodal data analysis is to evaluate if a whole modality is significantly associated with the outcome, given other modalities in the model. For instance, in multi-omics analysis, it is of interest to test if DNA methylation correlates with the phenotypic traits related to genetic disorders given gene expression level (Richardson et al., 2016). In multimodal neuroimaging analysis, it is of interest to evaluate if functional imaging quantification for hypometabolism associates with the diagnosis of Alzheimer's disease, given structural magnetic resonance imaging of brain atrophy measurement (Zhang et al., 2011). The challenge here is that even a single modality often contains many more variables than the sample size.

This is essentially a problem of testing a high-dimensional subvector of $\boldsymbol{\beta}^*$ in a high-dimensional regression model. Related testing problems have been extensively studied for a single modality data. For example, Zhang and Zhang (2014); Van de Geer et al. (2014); Javanmard and Montanari (2014) developed bias-corrected or de-sparsifying methods to test if a fixed-dimensional subvector of $\boldsymbol{\beta}^*$ in a high-dimensional linear or generalized linear model equals zero. In particular, under a general M-estimation framework, Ning and Liu (2017) proposed a decorrelated score test for the same problem, i.e., to test if a subvector $\boldsymbol{\beta}_S^* = \mathbf{0}$. They first showed that their score test statistic has a closed-form limiting distribution when the dimension of the subset $|S|$ is fixed. They then extended to the case where $\boldsymbol{\beta}_S^*$ can be any arbitrary subvector of $\boldsymbol{\beta}^*$ with $|S|$ diverging and even when $|S| > n$. Built on a pioneering work by Chernozhukov et al. (2013), they showed that the distribution of the supremum of the decorrelated score functions can be approximated by a multiplier bootstrap approach. Consequently, they employed bootstrap simulations to obtain the critical values of the limiting distribution to form the rejection region. Our test differs from Ning and Liu (2017). When $|S|$ diverges, the test of Ning and Liu (2017) no longer has a closed-form limiting distribution, and they had to resort to bootstrap for critical values. By contrast, we are able to obtain a closed-form limiting distribution for our test when $|S|$ diverges. This is due to that, instead of using the observed likelihood, we perform factor decomposition on $\mathbf{X}_m$ first, then use the factor model as a dimension reduction tool to reduce a high-dimensional test to a fixed-dimensional one. Our method does pay the extra price that we need to estimate the latent factors to plug into the likelihood function. However, as we show next, this extra cost can be well controlled. We also numerically compare with Ning and Liu (2017) in Section 7.1. We show that our test is as powerful, and often more powerful than the test of Ning and Liu (2017).

Formally, for our multimodal analysis, we aim at testing the following pair of hypotheses:

$$H_0 : \boldsymbol{\beta}_m^* = \mathbf{0} \quad \text{versus} \quad H_a : \boldsymbol{\beta}_m^* \neq \mathbf{0}.$$

We perform factor decomposition on the mth modality following (1). Then,

$$\mathbf{x}'\boldsymbol{\beta} = \mathbf{x}'_{-m}\boldsymbol{\beta}_{-m} + \mathbf{f}'_m\boldsymbol{\theta}_m + \mathbf{u}'_m\boldsymbol{\beta}_m,$$

where $\boldsymbol{\theta}_m = \boldsymbol{\Lambda}'_m\boldsymbol{\beta}_m$. The null hypothesis that $H_0 : \boldsymbol{\beta}_m^* = \mathbf{0}$ is then equivalent to $H_0 : \boldsymbol{\theta}_m^* = \mathbf{0}$, where $\boldsymbol{\theta}_m^* = \boldsymbol{\Lambda}'_m\boldsymbol{\beta}_m^*$. The difference, though, is that $\boldsymbol{\theta}_m^* \in \mathcal{R}^{K_m}$ is a low-dimensional vector, while $\boldsymbol{\beta}_m^* \in \mathcal{R}^{p_m}$ is high-dimensional. As such, the factor model plays the role of dimension reduction for our testing problem. Next we develop a factor-adjusted decorrelated score test, and show that the test statistic is still asymptotically efficient when the latent factors can be well estimated.

Following Ning and Liu (2017), and based on the Gaussian quasi-likelihood, we define the decorrelated score function as

$$\mathbf{S}(\boldsymbol{\beta}, \boldsymbol{\theta}_m) = \frac{1}{n\sigma_\epsilon^2}\sum_{i=1}^n (y_i - \mathbf{f}'_{i,m}\boldsymbol{\theta}_m - \mathbf{z}'_i\boldsymbol{\beta})(\mathbf{f}_{i,m} - \mathbf{V}'\mathbf{z}_i),$$

where $\mathbf{z}_i = (\mathbf{x}'_{i,-m}, \mathbf{u}'_{i,m})' \in \mathcal{R}^p$, and $\mathbf{V} \in \mathcal{R}^{p\times K_m}$. Unlike Ning and Liu (2017), we need to estimate the latent factors in our decorrelated score function $\mathbf{S}(\boldsymbol{\beta}, \boldsymbol{\theta}_m)$. That is, under the null hypothesis, we estimate $\mathbf{S}(\boldsymbol{\beta}, \boldsymbol{\theta}_m)$ by

$$\widehat{\mathbf{S}}(\widehat{\boldsymbol{\beta}}_{-m}, \mathbf{0}) = \frac{1}{n\widehat{\sigma}_\epsilon^2}\sum_{i=1}^n (y_i - \mathbf{x}'_{i,-m}\widehat{\boldsymbol{\beta}}_{-m})(\widehat{\mathbf{f}}_{i,m} - \widehat{\mathbf{W}}'\mathbf{x}_{i,-m}),$$

where $\widehat{\mathbf{f}}'_{i,m}$ is the ith row of the estimated latent factor matrix $\widehat{\mathbf{F}}_m$, $\widehat{\sigma}_\epsilon^2$ is any consistent estimator of σ_ϵ^2 that satisfies Condition 5 below, $\widehat{\boldsymbol{\beta}}_{-m} \in \mathcal{R}^{p_{-m}}$ and $\widehat{\mathbf{W}} \in \mathcal{R}^{p_{-m}\times K_m}$ are obtained by solving the following optimizations,

$$\begin{aligned}
(\widehat{\boldsymbol{\beta}}_{-m}, \widehat{\boldsymbol{\theta}}_m) &= \underset{(\boldsymbol{\beta}_{-m}, \boldsymbol{\theta}_m)}{\operatorname{argmin}} \frac{1}{2n}\sum_{i=1}^n \left(y_i - \mathbf{x}'_{i,-m}\boldsymbol{\beta}_{-m} - \widehat{\mathbf{f}}'_{i,m}\boldsymbol{\theta}_m\right)^2 + \lambda_1\|\boldsymbol{\beta}_{-m}\|_1, \\
\widehat{\mathbf{W}} &= \operatorname{argmin}\|\mathbf{W}\|_1, \text{ such that } \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{x}_{i,-m}\left(\widehat{\mathbf{f}}'_{i,m} - \mathbf{x}'_{i,-m}\mathbf{W}\right)\right\|_\infty \le \lambda_2.
\end{aligned}$$

The solution $\widehat{\mathbf{W}}$ is an estimator of $\mathbf{W}^* = \mathrm{E}(\mathbf{x}_{i,-m}^{\otimes 2})^{-1}\mathrm{E}(\mathbf{x}_{i,-m}\mathbf{f}'_{i,m}) \in \mathcal{R}^{p_{-m}\times K_m}$, where $p_{-m} = p - p_m$. We assume $\mathbf{W}^*$ is column-wise sparse; see Condition 4.

By the sandwich formula, the information matrix is

$$\mathbf{I}^*_{\boldsymbol{\theta}_m|\boldsymbol{\beta}_{-m}} = \sigma_\epsilon^{-2}\left\{\mathrm{E}(\mathbf{f}_{i,m}^{\otimes 2}) - \mathrm{E}(\mathbf{f}_{i,m}\mathbf{x}'_{i,-m})\mathrm{E}(\mathbf{x}_{i,-m}^{\otimes 2})^{-1}\mathrm{E}(\mathbf{x}_{i,-m}\mathbf{f}'_{i,m})\right\} \in \mathcal{R}^{K_m\times K_m},$$

which can be estimated by

$$\widehat{\mathbf{I}}_{\boldsymbol{\theta}_m|\boldsymbol{\beta}_{-m}} = \widehat{\sigma}_\epsilon^{-2}\left\{\frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{f}}_{i,m}^{\otimes 2} - \widehat{\mathbf{W}}'\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_{i,-m}\widehat{\mathbf{f}}'_{i,m}\right)\right\}.$$

Then, our test statistic is given by

$$T_s = \sqrt{n}\,\widehat{\mathbf{I}}^{-1/2}_{\boldsymbol{\theta}_m|\boldsymbol{\beta}_{-m}}\widehat{\mathbf{S}}(\widehat{\boldsymbol{\beta}}_{-m}, \mathbf{0}).$$

We next show that, under the null hypothesis, the asymptotic distribution of T_s is $N(\mathbf{0}, \mathbf{I}_{K_m})$. In other words, T_s is asymptotically efficient. We first begin with a set of conditions.

Condition 1. For all $m \in [M]$, suppose $\{(\mathbf{f}'_{i,m}, \mathbf{u}'_{i,m})'\}_{i=1}^n$ are i.i.d. uncorrelated sub-Gaussian random vectors with zero mean. That is, $\mathrm{E}(\mathbf{f}_{i,m}) = \mathbf{0}$, $\mathrm{E}(\mathbf{u}_{i,m}) = \mathbf{0}$, and $\mathrm{E}(\mathbf{f}_{i,m}\mathbf{u}'_{i,m}) = \mathbf{0}$. Moreover, $\mathrm{E}\{\exp(t\boldsymbol{\alpha}'\mathbf{f}_{i,m})\} \le \exp(-C\|\boldsymbol{\alpha}\|_2^2 t^2/2)$, and $\mathrm{E}\{\exp(t\boldsymbol{\alpha}'\mathbf{u}_{i,m})\} \le \exp(-C\|\boldsymbol{\alpha}\|_2^2 t^2/2)$, for some constant C. In addition, for all $k \in [K_m]$, $\mathbf{x}'_{i,-m}\mathbf{w}_k$ are i.i.d. sub-Gaussian such that $\mathrm{E}\{\exp(t\mathbf{x}'_{i,-m}\mathbf{w}_k)\} \le \exp(-Ct^2/2)$. Additionally, $\{\epsilon_i\}_{i=1}^n$ are i.i.d. sub-Gaussian with zero mean, and ϵ_i is uncorrelated with $(\mathbf{f}'_{i,m}, \mathbf{u}'_{i,m})'$ for all $m \in [M]$.

Condition 2. For all $m \in [M]$, suppose $0 < c \le \lambda_{\min}(\boldsymbol{\Lambda}'_m\boldsymbol{\Lambda}_m/p_m) \le \lambda_{\max}(\boldsymbol{\Lambda}'_m\boldsymbol{\Lambda}_m/p_m) \le C < \infty$, for some positive constants c and C.

Condition 3. For all $m \in [M]$, $s, t \in [p_m]$, $i, j \in [n]$, suppose $\mathrm{E}[p_m^{-1/2}\{\mathbf{u}'_{i,m}\mathbf{u}_{j,m} - \mathrm{E}(\mathbf{u}'_{i,m}\mathbf{u}_{j,m})\}^4] \le C$, and $\mathrm{E}\|p_m^{-1/2}\boldsymbol{\Lambda}'_m\mathbf{u}_{i,m}\|_2^4 \le C$. Moreover, $\|\boldsymbol{\Lambda}_m\|_\infty \le C$, $\lambda_{\min}(\boldsymbol{\Sigma}_{u_m}) > c$, $\|\boldsymbol{\Sigma}_{u_m}\|_1 \le C$, and $\min_{s,t\in[p_m]}\mathrm{Var}(u_{i,m_s}u_{i,m_t}) > c$.

Condition 4. Let $\mathbf{w}^*_k$ be the kth column of $\mathbf{W}^*$ and $s^*_w = \max_{k\in[K_m]}|\mathrm{supp}(\mathbf{w}^*_k)|$. Suppose $s^*_w \log(p_{-m})\{1\vee(n^{1/4}/\sqrt{p_m})\} = o(n^{1/2})$, and $[s^*_{-m}\{\sqrt{(\log p_{-m})/n} + 1/\sqrt{p_m}\}]\cdot\sqrt{\log(p_{-m})}\{1\vee(n^{1/4}/\sqrt{p_m})\} = o(1)$.

Condition 5. Suppose $\widehat{\sigma}_\epsilon^2 = \sigma_\epsilon^2 + o_P(1)$.

Condition 1 is a typical sub-Gaussian assumption for high-dimensional problems. Condition 2 is the pervasive condition, and is common in factor analysis (Fan et al., 2013). It requires that the latent factors affect a large number of variables. This is reasonable for a variety of multimodal data. For instance, in multi-omics data, some genetic factors are believed to impact both gene expression and DNA methylation, and in multimodal neuroimaging, some neurological factors affect both brain structures and functions. Condition 3 imposes some technical requirements on the loading matrix and idiosyncratic component. Together, Conditions 2 and 3 ensure that $\mathbf{f}_{i,m}$ and $\mathbf{u}_{i,m}$ can be consistently estimated by the PCA method (Fan et al., 2013). Condition 4 is a sparsity condition on $\mathbf{W}^*$ and $\boldsymbol{\beta}^*_{-m}$, which requires s^*_w and s^*_{-m} to be much smaller than n. Under such a condition, $\mathbf{W}^*$ and $\boldsymbol{\beta}^*_{-m}$ can both be consistently estimated, even if the latent factors are unknown.Condition 5 ensures the estimator of σ^2_ϵ is consistent.

We next obtain a closed-form limiting distribution for the test statistic T_s.

Theorem 1. *Suppose Conditions 1–5 hold. Suppose* $\lambda_1 \asymp \sqrt{(\log p_{-m})/n} + 1/\sqrt{p_m}$, *and* $\lambda_2 \asymp \sqrt{(\log p_{-m})/n}\{1 \vee (n^{1/4}/\sqrt{p_m})\}$. *Then, under* $H_0 : \boldsymbol{\beta}^*_m = \mathbf{0}$, *it holds that*

$$T_s \xrightarrow{D} N(\mathbf{0}, \mathbf{I}_{K_m}).$$

By Theorem 1, we reject the null hypothesis if $n\{\widehat{\mathbf{S}}(\widehat{\boldsymbol{\beta}}_{-m}, \mathbf{0})\}' \widehat{\mathbf{I}}^{-1}_{\boldsymbol{\theta}_m|\boldsymbol{\beta}_{-m}} \widehat{\mathbf{S}}(\widehat{\boldsymbol{\beta}}_{-m}, \mathbf{0}) > \chi^2_\alpha(K_m)$, where $\chi^2_\alpha(K_m)$ is the α-upper quantile of the χ^2-distribution with K_m degrees of freedom.

We next explicitly discuss the benefit and the extra cost of our factor-based test when compared with Ning and Liu (2017). The main difference is that, through latent factors, we obtain a closed-form limiting distribution and do not have to resort to bootstrap. The price we pay mainly lies in Condition 4 and the choices of λ_1 and λ_2. Actually, the extra term $1/\sqrt{p_m}$ appearing in both Condition 4 and λ_1, λ_2 reflects the estimation error caused by using $\widehat{\mathbf{f}}_{i,m}$ to estimate $\boldsymbol{\beta}^*_{-m}$ in $\mathbf{S}(\boldsymbol{\beta}, \boldsymbol{\theta}_m)$. The term $n^{1/4}/\sqrt{p_m}$ is due to the same reason for estimating $\mathbf{w}^*_k$. Therefore, the choices of the tuning parameters λ_1 and λ_2 need to be adjusted accordingly, by taking into account such extra estimation errors.

We further consider three scenarios. First, when $p_m \gg n$, both $1/\sqrt{p_m}$ and $n^{1/4}/\sqrt{p_m}$ are dominated by $\sqrt{(\log p_{-m})/n}$. Therefore, the estimation errors of $\widehat{\boldsymbol{\beta}}_{-m}$ and $\widehat{\mathbf{W}}$ reach the optimal oracle rate, i.e. the best rate as if the latent factors were known.In this case, using the factor estimators actually does not pay any extra cost. The reason is that many variables are used to estimate the latent factors, and its estimation error is so small that it would not affect the inference on $\boldsymbol{\beta}^*_m$. Second, when $p_m = o(n)$, the estimation errors of

$\widehat{\boldsymbol{\beta}}_{-m}$ and $\widehat{\mathbf{W}}$ would be greater than the optimal rate. However, the central limit theorem still holds, given proper choices of λ_1 and λ_2, and more stringent sparsity conditions on $\boldsymbol{\beta}^*_{-m}$ and $\mathbf{W}^*$ in Condition 4. Third, in a special case where variables in all modalities are driven by exactly the same latent factors, even we perform the hypothesis test on the mth modality, we could use variables from all different modalities to estimate the latent factors. Then, the terms $1/\sqrt{p_m}$ and $n^{1/4}/\sqrt{p_m}$ become $1/\sqrt{p}$ and $n^{1/4}/\sqrt{p}$, respectively, which are naturally dominated by $\sqrt{(\log p_{-m})/n}$. In this case, the optimal oracle rate is again attained. Such a result can be viewed as a blessing of the dimensionality for the factor model. In summary, our method is most suitable for testing the significance of a modality containing many variables, or for multimodal data with a large number of variables driven by some common latent factors.

5 Hypothesis test of a linear combination of predictors of one or more modalities

Another important question in multimodal data analysis is to test if a linear combination of predictors, within the same modality or across different modalities, is significantly correlated with the response. This is because multimodal data often measures different aspects of related quantities. For instance, in multi-omics studies, expression data measures how genes are expressed, methylation data measures how DNA molecules are methylated, and both data may be related to the same set of genes. In multimodal neuroimaging analysis, brain structures, functions, and chemical constituents of the same brain regions are often measured simultaneously. As such, it is of great scientific interest to test if various measurements on a particular gene or brain region are associated with the outcome.

Shi et al. (2019) considered a similar testing problem in a high-dimensional generalized linear model for a single modality data. They derived the corresponding partially penalized likelihood ratio test, score test and Wald test, and showed that the three tests are asymptotically equivalent. They allowed the dimension of the model to grow with the sample size, as long as the dimension of the subvector being tested and the number of linear combinations are smaller than the sample size. Our method differs from Shi et al. (2019) in several ways. Shi et al. (2019) treated the design matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)$ as fixed, while we treat $\mathbf{X}$ as i.i.d. random realizations from some distributions. More importantly, we do not directly use the observed $\mathbf{X}$, but instead perform a factor decomposition and use the decorrelated idiosyncratic components as the pseudo design matrix. We explicitly show in Theorem 2 that such a factor-adjusted step leads to less stringent conditions to reach the variable selection and estimation consistency. Moreover, since variable selection consistency is needed to correctly calculate the variance of the test statistic, as shown in Theorem 3, our method also requires less stringent conditions to establish the limiting distribution of the test statistic. We further numerically compare with Shi et al. (2019) in Section 7.2, and show that the improved variable selection by the factor-adjusted step in turn improves the power of the test. Finally, our model concerns with data with multiple modalities, instead of a single modality as in Shi et al. (2019). High correlations are commonly observed in multimodal data, and as such the factor-adjusted decorrelation step becomes essential.

Formally, we consider testing the following pair of hypotheses:

$$H_0 : \mathbf{A}\boldsymbol{\beta}_T^* = \mathbf{b} \quad \text{versus} \quad H_a : \mathbf{A}\boldsymbol{\beta}_T^* \neq \mathbf{b},$$

where $\mathbf{A} \in \mathcal{R}^{r \times t}$, $\mathbf{b} \in \mathcal{R}^r$, $\boldsymbol{\beta}_T^* \in \mathcal{R}^t$ is a subvector of $\boldsymbol{\beta}^*$, and $T \subset [d]$ is a low-dimensional index set with $|T| = t < n$. This simultaneously tests r linear combinations of $\boldsymbol{\beta}_T^*$, with $r < n$. We next develop a factor-adjusted Wald test.

To construct the test statistic, we first consider a penalized least squares problem,

$$(\widehat{\boldsymbol{\gamma}}_a, \widehat{\boldsymbol{\beta}}_a) = \underset{(\boldsymbol{\gamma},\boldsymbol{\beta})}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \widehat{\mathbf{F}}_i' \boldsymbol{\gamma} - \widehat{\mathbf{U}}_i' \boldsymbol{\beta} \right)^2 + \lambda_a \sum_{j \notin T} p(|\beta_j|). \tag{5}$$

This is essentially the same as (4), except that, instead of penalizing all variables in $\boldsymbol{\beta}$, we only penalize β_j for $j \notin T$ this time. Such a penalization is to avoid introducing bias when estimating β_j^* for $j \in T$, and is similar as that in Shi et al. (2019).

Given $\widehat{\boldsymbol{\beta}}_a$, our factor-adjusted Wald test statistic is given by

$$T_w = (\mathbf{A}\widehat{\boldsymbol{\beta}}_{a,T} - \mathbf{b})'(\mathbf{A}\widehat{\boldsymbol{\Omega}}_T\mathbf{A}')^{-1}(\mathbf{A}\widehat{\boldsymbol{\beta}}_{a,T} - \mathbf{b})/\widehat{\sigma}_\epsilon^2,$$

where $\widehat{\boldsymbol{\beta}}_{a,T}$ is the sub-vector of $\widehat{\boldsymbol{\beta}}_a$ with indices in T, $\widehat{\boldsymbol{\Omega}}_T$ is the first T rows and columns of

$$\widehat{\boldsymbol{\Omega}}_{T \cup \widehat{S}_a} = n \begin{pmatrix} \widehat{\mathbf{U}}_T' \widehat{\mathbf{U}}_T & \widehat{\mathbf{U}}_T' \widehat{\mathbf{U}}_{\widehat{S}_a} \\ \widehat{\mathbf{U}}_{\widehat{S}_a}' \widehat{\mathbf{U}}_T & \widehat{\mathbf{U}}_{\widehat{S}_a}' \widehat{\mathbf{U}}_{\widehat{S}_a} \end{pmatrix}^{-1},$$

$\widehat{S}_a = \{j \in T^c : \widehat{\beta}_{a,j} \neq 0\}$, and $\widehat{\sigma}_\epsilon^2$ is any consistent estimator of σ_ϵ^2. In this test statistic, $\widehat{S}_a$ plays a critical role in calculating the variance of $\mathbf{A}\widehat{\boldsymbol{\beta}}_{a,T} - \mathbf{b}$. In fact, $\widehat{S}_a$ needs to be consistent to $S_a = \{j \in T^c : \beta_j^* \neq 0\}$ in order for the variance to be valid. Such a consistency is guaranteed by Theorem 2.

We next present a set of regularity conditions.

Condition 6. Suppose $c \leq \lambda_{\min}\{\mathrm{E}(\mathbf{u}^{\otimes 2})\} \leq \lambda_{\max}\{\mathrm{E}(\mathbf{u}^{\otimes 2})\} \leq C$ for some postive constants c and C.

Condition 7. Suppose $\|\mathrm{E}(\mathbf{u}_{T\cup S_a}^{\otimes 2})^{-1}\|_{L_\infty} \leq C$.

Condition 8. Suppose $\|\mathrm{E}(\mathbf{u}_{(T\cup S_a)^c}\mathbf{u}'_{T\cup S_a})\{\mathrm{E}(\mathbf{u}_{T\cup S_a}^{\otimes 2})^{-1}\}\|_{L_\infty} \leq C$.

Condition 9. Suppose $d_n = \min\{|\beta_j^*| : \beta_j^* \neq 0\}/2 \gg \lambda_a \gg \delta_n$, where $\delta_n = \sqrt{(\log p)/n}\{1 \vee (n^{1/4}/\sqrt{p_{\min}})\}$, $p_{\min} = \min_{m\in[M]} p_m$, and $\lambda_a \dot{p}(d_n) = o(\delta_n)$, where $\dot{p}$ is the first derivative.

We first note that Conditions 6–8 are imposed on $\mathbf{u}$, instead of on $\mathbf{x}$. Since $\mathbf{u}$ can be viewed as the residual of $\mathbf{x}$ after the latent factors are removed, the correlations among the variables in $\mathbf{u}$ are much weaker than those in $\mathbf{x}$. In particular, Conditions 6 and 7 are needed to avoid singularity of $\mathrm{E}(\mathbf{u}^{\otimes 2})$ and $\mathrm{E}(\mathbf{u}_{T\cup S_a}^{\otimes 2})$. Condition 8 is the well-known irrepresentable condition, which is necessary for establishing the variable selection consistency. Shi et al. (2019) required such a condition to hold for the Gram matrix $\mathbf{X}'\mathbf{X}$, which essentially requires the correlations among $\mathbf{X}$ must be small. This condition hardly holds for multimodal data. By contrast, we only impose such a condition on $\mathrm{E}(\mathbf{u}^{\otimes 2})$, which requires the idiosyncratic components not to be highly correlated. This condition is well accepted in the factor model literature. Indeed, when an exact factor model is assumed, $\mathrm{E}(\mathbf{u}^{\otimes 2})$ is a diagonal matrix, then Condition 6 naturally holds.

We now establish the variable selection and estimation consistency of the estimator $\widehat{\boldsymbol{\beta}}_a$ in (5), which is essential for deriving the asymptotic distribution of T_w.

Theorem 2. *Suppose Conditions 1–3 and 6–9 hold. Then there exists a solution $(\widehat{\boldsymbol{\beta}}_a, \widehat{\boldsymbol{\gamma}}_a)$ of (5) such that, with probability tending to 1, the following results hold:*

(a) (sign consistency) $\mathrm{sign}(\widehat{\boldsymbol{\beta}}_a) = \mathrm{sign}(\boldsymbol{\beta}^*)$*;*

(b) (L_∞ consistency) $\|\widehat{\boldsymbol{\beta}}_{a,T\cup S_a} - \boldsymbol{\beta}^*_{T\cup S_a}\|_\infty = O_P(\delta_n)$*;*

(c) (L_2 consistency) $\|\widehat{\boldsymbol{\beta}}_{a,T\cup S_a} - \boldsymbol{\beta}^*_{T\cup S_a}\|_2 = O_P\left(\sqrt{t+s_a}\delta_n\right)$*, where* $s_a = |S_a|$*;*

(d) (asymptotic expansion) $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{a,T\cup S_a} - \boldsymbol{\beta}^*_{T\cup S_a}) = n^{-1/2}\mathbf{K}_n^{-1}\mathbf{U}'_{T\cup S_a}\boldsymbol{\epsilon} + o_P(1)$*, where* $\mathbf{K}_n = (1/n)\mathbf{U}'_{T\cup S_a}\mathbf{U}_{T\cup S_a}$*, if we further have that* $p_{\min} \gg n^{3/2}$*, and* $\sqrt{n}\lambda_a \dot{p}(d_n) = o(1)$*.*

We again explicitly examine the benefit and the extra cost of our factor-based test compared with Shi et al. (2019). The main difference is that we obtain the variable selection and estimation consistency under much weaker conditions than Shi et al. (2019). The price we pay lies in δ_n, which reflects the convergency rates in (b) and (c) of Theorem 2. Particularly, the component $n^{1/4}/\sqrt{p_{\min}}$ in δ_n is due to the factor estimation. We consider two scenarios. First, when all data modalities have a large number of variables, i.e. $p_{\min} \gg n^{1/2}$, then $\delta_n = \sqrt{(\log p)/n}$, which makes the convergence rates in (b) and (c) to be minimax optimal. This is because when there are enough variables to estimate the latent factors well, the extra factor estimation error becomes so small that it would not affect the estimation error on $\widehat{\boldsymbol{\beta}}_a$. Second, when one modality has only a small number of variables, i.e. $p_m = o(n^{1/4})$ for some $m \in [M]$, estimating the latent factors in that modality becomes challenging, and the resulting estimation error would slow the convergence of $\widehat{\boldsymbol{\beta}}_a$. In this case, one possible alternative solution is to skip factor decomposition for that particular modality, but directly use $\mathbf{X}_m$ in (5) and solve

$$(\widehat{\boldsymbol{\gamma}}_a, \widehat{\boldsymbol{\beta}}_a) = \underset{(\boldsymbol{\gamma},\boldsymbol{\beta})}{\operatorname{argmin}} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \mathbf{X}'_{i,m}\boldsymbol{\beta}_m - \widehat{\mathbf{F}}'_{i,-m}\boldsymbol{\gamma} - \widehat{\mathbf{U}}'_{i,-m}\boldsymbol{\beta}_{-m}\right)^2 + \lambda_a \sum_{j\notin T} p(|\beta_j|),$$

where $\mathbf{X}'_{i,m}$, $\widehat{\mathbf{F}}'_{i,-m}$ and $\widehat{\mathbf{U}}'_{i,-m}$ denote the ith row of $\mathbf{X}_{-m}$, $\widehat{\mathbf{F}}_{-m}$ and $\widehat{\mathbf{U}}_{-m}$, respectively. Finally, we note that the variable selection and estimation consistency of $\widehat{\boldsymbol{\beta}}$ in (4) is directly implied by Theorem 2 if we treat T as the empty set.

Next, we study the asymptotic distribution of our test statistic T_w, and show that it can be uniformly approximated by a χ^2-distribution under both H_0 and H_a. We need two more regularity conditions.

Condition 10. Suppose $\|\mathbf{h}_n\|_2 = O_P(\sqrt{r/n})$, and $\lambda_{\max}\{(\mathbf{A}\mathbf{A}')^{-1}\} \le C$ for some constant C, where $\mathbf{h}_n = \mathbf{A}\boldsymbol{\beta}^*_T - \mathbf{b}$.

Condition 11. Suppose $r^{1/4}n^{-1/2}\mathrm{E}|\mathbf{u}'_{T\cup S_a}\boldsymbol{\Sigma}^{-1}_{u,T\cup S_a}\mathbf{u}_{T\cup S_a}|^{3/2} \to 0$, where $\boldsymbol{\Sigma}^{-1}_{u,T\cup S_a}$ is the inverse of the submatrix of $\boldsymbol{\Sigma}_u$ with rows and columns in $T\cup S_a$.

Condition 10 regulates the local alternative $\mathbf{h}_n$ and avoids singularity of $\mathbf{A}\mathbf{A}'$. Condition 11 is a Lyapunov condition to ensure the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{a,T\cup S_a}$, which is the key to establish the χ^2-approximation.

Theorem 3. *Suppose the conditions of Theorem 2 and Conditions 10 and 11 hold,* $p_{\min} \gg n^{3/2}$, $\sqrt{n}\lambda_a \dot{p}(d_n) = o(1)$, *and* $t + s_a = o(n^{1/3})$. *Then it holds that*

$$\sup_x \left| \Pr(T_w \leq x) - \Pr\{\chi^2(r, \nu_n) \leq x\} \right| \to 0,$$

where $\nu_n = n\mathbf{h}_n'(\mathbf{A}\boldsymbol{\Omega}_T\mathbf{A}')^{-1}\mathbf{h}_n/\sigma_\epsilon^2$, $\boldsymbol{\Omega}_T$ *is the the submatrix of* $\boldsymbol{\Sigma}_{u,T\cup S_a}^{-1}$ *with rows and columns in* T.

By Theorem 3, we reject $H_0 : \mathbf{A}\boldsymbol{\beta}_T^* = \mathbf{b}$ if $T_w > \chi_\alpha^2(r)$, where $\chi_\alpha^2(r)$ is the α-upper quantile of the χ^2-distribution with r degrees of freedom. The limiting distribution we establish in Theorem 3 is the same as the classical Wald test result for a low-dimensional linear regression model (Shi et al., 2019).

We also remark that, the requirement of $p_{\min} \gg n^{3/2}$ in Theorem 3 ensures that the latent factors in each modality can be well estimated. Therefore, the extra factor estimation error would not affect the limiting distribution of T_w. This condition is more stringent than that of $p_{\min} \gg n^{1/2}$, which guarantees the minimax optimal rate of estimation in Theorem 2. This is because hypothesis testing is a more challenging task than estimation.

6 Quantification of contribution of a single modality

In addition to testing the significance of a whole data modality, it is of equal interest to quantify the amount of contribution of a modality conditioning on other data modalities in the regression model. As an example, in heritability analysis, the goal is to evaluate the contribution of genetic effects to the phenotype in addition to the environmental effects (Lynch et al., 1998). Motivated by the proportion of the response variance explained in the classical linear regression, we propose a measure of the contribution of a single data modality in our integrative factor regression model.

Let $\mathbf{x}_{-m} \in \mathcal{R}^{p-p_m}$ denote the subvector of $\mathbf{x} \in \mathcal{R}^p$ excluding $\mathbf{x}_m \in \mathcal{R}^{p_m}$, and $\mathbf{X}_{-m} \in \mathcal{R}^{n\times(p-p_m)}$ denote the submatrix of $\mathbf{X} \in \mathcal{R}^{n\times p}$ excluding $\mathbf{X}_m \in \mathcal{R}^{n\times p_m}$. To evaluate the contribution of $\mathbf{x}_m$, our key idea is to compare the goodness-of-fit of regressing y on $\mathbf{x}$ to that of regressing y on $\mathbf{x}_{-m}$. Under model (2) along with (1), we define

$$\sigma^2_{m|-m} = \mathrm{Var}(\mathbf{x}_m' \boldsymbol{\beta}^*_m | \mathbf{x}_{-m}).$$

The next proposition shows that $\sigma^2_{m|-m}$ quantifies the contribution of $\mathbf{x}_m$ in terms of the goodness-of-fit, or equivalently, additional variance of the response explained. We first consider the scenario where $p < n$, which helps simplify the presentation. Later we discuss the more general scenario where $p > n$. To derive a closed-form expression of $\sigma^2_{m|-m}$, we further assume $\mathbf{x}_m$ and $\mathbf{x}_{-m}$ share some common factors and have decompositions as $\mathbf{x}_m = \boldsymbol{\Lambda}_m \mathbf{f} + \mathbf{u}_m$ and $\mathbf{x}_{-m} = \boldsymbol{\Lambda}_{-m} \mathbf{f} + \mathbf{u}_{-m}$, where $\mathbf{f} \in \mathcal{R}^K$.

Proposition 1. *Suppose $\mathbf{x}$ follows a multivariate normal distribution and $p < n$. Let $\widehat{\mathbf{Y}}$ and $\widehat{\mathbf{Y}}_{-m}$ denote the predicted response values by regressing y on $\mathbf{x}$, and regressing y on $\mathbf{x}_{-m}$, respectively. Let σ^2_y denote the variance of y. Then the following results hold:*

(a) $\sigma^2_{m|-m} = \mathrm{E}\|\mathbf{Y} - \widehat{\mathbf{Y}}_{-m}\|_2^2/(n - p_{-m}) - \sigma^2_\epsilon$;

(b) $\sigma^2_{m|-m} = \sigma^2_y(r^2 - r^2_{-m})$, *where* $r^2 = 1 - \mathrm{E}\|\mathbf{Y} - \widehat{\mathbf{Y}}\|_2^2/\{(n-p)\sigma^2_y\}$, *and* $r^2_{-m} = 1 - \mathrm{E}\|\mathbf{Y} - \widehat{\mathbf{Y}}_{-m}\|_2^2/\{(n - p_{-m})\sigma^2_y\}$;

(c) $\sigma^2_{m|-m} = \boldsymbol{\beta}^{*\prime}_m \left\{ \boldsymbol{\Lambda}_m (\mathbf{I}_K + \boldsymbol{\Lambda}'_{-m} \boldsymbol{\Sigma}^{-1}_{u_{-m}} \boldsymbol{\Lambda}_{-m})^{-1} \boldsymbol{\Lambda}'_m + \boldsymbol{\Sigma}_{u_m} \right\} \boldsymbol{\beta}^*_m$.

The results in Proposition 1(a) and (b) give two ways to understand why $\sigma^2_{m|-m}$ can be used to quantify the contribution of the modality $\mathbf{x}_m$. The result in Proposition 1(c) points a way to estimate $\sigma^2_{m|-m}$ given the data.

First, by Proposition 1(a), when regressing y using all but the mth modality, we have

$$\mathrm{E}\|\mathbf{Y} - \widehat{\mathbf{Y}}_{-m}\|_2^2 = (n - p_{-m})(\sigma^2_\epsilon + \sigma^2_{m|-m}).$$

On the other hand, when regressing y on all data modalities, we have $\mathrm{E}\|\mathbf{Y} - \widehat{\mathbf{Y}}\|_2^2 = (n-p)\sigma^2_\epsilon$. Therefore, from a goodness-of-fit perspective, ignoring $\mathbf{x}_m$ leads to a "worsened" prediction by an amount of $\sigma^2_{m|-m}$.

Second, for Proposition 1(b), recall in the classical linear regression model, the adjusted R^2 measures the percentage of total variation in the response that has been explained by the predictors, and is defined as

$$R^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)},$$

where RSS and TSS are the residual sum of squares and total sum of squares, respectively. Then, r^2 in Proposition 1(b) can be viewed as an "expected" percentage of total variation in the response explained, in that,

$$r^2 = 1 - \frac{\mathrm{E}(RSS)/(n-p)}{\mathrm{E}(TSS)/(n-1)} = 1 - \frac{\mathrm{E}\|\mathbf{Y} - \widehat{\mathbf{Y}}\|_2^2}{(n-p)\sigma_y^2}.$$

As we show in the proof of Proposition 1, when using all but the mth modality, the "expected" percentage of total variation in the response explained is $r_{-m}^2 = 1 - (\sigma_\epsilon^2 + \sigma_{m|-m}^2)/\sigma_y^2$, where $\sigma_\epsilon^2 = \mathrm{E}\|\mathbf{Y} - \widehat{\mathbf{Y}}\|_2^2/(n-p)$. On the other hand, when using all data modalities, the "expected" percentage of total variation in the response explained is $r^2 = 1 - \sigma_\epsilon^2/\sigma_y^2$. Therefore, using the mth modality improves the "expected" percentage of total variation in the response explained by an amount of $\sigma_{m|-m}^2/\sigma_y^2$.

We have so far justified $\sigma_{m|-m}^2$ in the low-dimensional setting where $p < n$. In the high-dimensional setting where $p > n$, we can still use this same measure to quantify the contribution of an individual modality. The above justifications continue to hold, by replacing $(n-p)$ with $(n - df)$, where df is the degrees of freedom of the true full model. However, an appealing feature of $\sigma_{m|-m}^2$ is that we do not need to estimate the degrees of freedom df of the full regression model, which can be difficult in a high-dimensional setting.

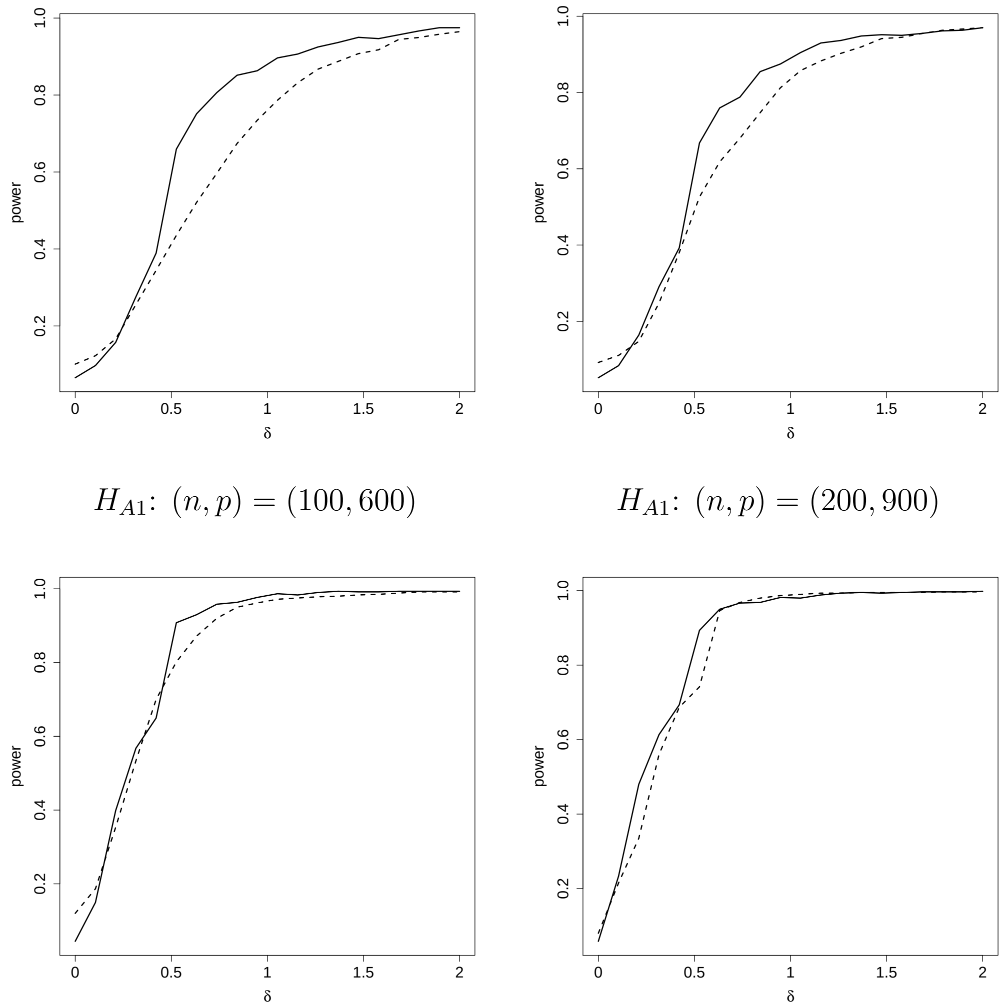
Finally, Proposition 1(c) gives a closed-form expression of $\sigma_{m|-m}^2$, and a way to estimate $\sigma_{m|-m}^2$ given the data. Each component in this expression can be consistently estimated. First, we obtain $\widehat{\mathbf{F}}$, by performing PCA on the concatenated data matrices $\mathbf{X}$, since $\sigma_{m|-m}^2$ leverages on the shared common factors between $\mathbf{x}_m$ and $\mathbf{x}_{-m}$. Then, for each $m \in [M]$, we estimate $\mathbf{\Lambda}_m$ by $\widehat{\mathbf{\Lambda}}_m = (1/n)\mathbf{X}_m'\widehat{\mathbf{F}}$, and obtain $\widehat{\mathbf{U}}_m = \mathbf{X}_m - \widehat{\mathbf{F}}\widehat{\mathbf{\Lambda}}_m'$. Next we solve for $\widehat{\boldsymbol{\beta}}$ following (4). By Theorem 2, $\widehat{\boldsymbol{\beta}}_m$ is a consistent estimator of $\boldsymbol{\beta}_m^*$. By Fan et al. (2013), $\widehat{\mathbf{\Lambda}}_m$ and $\widehat{\mathbf{\Lambda}}_{-m}$ are two consistent estimators as well. For $\mathbf{\Sigma}_u$, we consider a thresholded estimator $\widehat{\mathbf{\Sigma}}_u$, whose (i,j)th element is $\widehat{\sigma}_{u,ij}^2 = s(n^{-1}\sum_{\ell=1}^n \widehat{U}_{\ell i}\widehat{U}_{\ell j}, \omega)$, $s(x,\omega)$ is a thresholding function, ω is the threshold, and $\widehat{\mathbf{U}} = (\widehat{\mathbf{U}}_1, \ldots, \widehat{\mathbf{U}}_M)$. By Theorem 3.1 of Fan et al. (2013), as long as ω is chosen properly, $\widehat{\mathbf{\Sigma}}_u$ is a consistent estimator of $\mathbf{\Sigma}_u$. Plugging all these estimators into Proposition 1(c) gives a consistent estimator of $\sigma_{m|-m}^2$.

We make two additional remarks about $\sigma^2_{m|-m}$. First, the closed-form expression of $\sigma^2_{m|-m}$ utilizes the factors commonly shared by $\mathbf{x}_m$ and $\mathbf{x}_{-m}$. Indeed, such factors determine the correlations between $\mathbf{x}_m$ and $\mathbf{x}_{-m}$. When no such common factors exist, $\mathbf{x}_m$ and $\mathbf{x}_{-m}$ are uncorrelated. In that case, $\mathrm{Var}(\mathbf{x}_m'\boldsymbol{\beta}_m^*|\mathbf{x}_{-m}) = \mathrm{Var}(\mathbf{x}_m'\boldsymbol{\beta}_m^*) = \boldsymbol{\beta}_m^{*\prime}\boldsymbol{\Sigma}_{x_m}\boldsymbol{\beta}_m^* = \boldsymbol{\beta}_m^{*\prime}\boldsymbol{\Sigma}_{u_m}\boldsymbol{\beta}_m^*$. Therefore, the closed-form expression in Proposition 1(c) can be viewed as a more general form of this special case by taking the correlations between $\mathbf{x}_m$ and $\mathbf{x}_{-m}$ into account. Second, the computation of $\sigma^2_{m|-m}$ only requires to invert a sparse high-dimensional matrix $\boldsymbol{\Sigma}_{u_{-m}}$ and a low-dimensional matrix $\mathbf{I}_K + \boldsymbol{\Lambda}_{-m}'\boldsymbol{\Sigma}_{u_{-m}}^{-1}\boldsymbol{\Lambda}_{-m}$. If an exact factor model is further adopted such that $\boldsymbol{\Sigma}_u$ becomes a diagonal matrix, $\sigma^2_{m|-m}$ can be easily computed, as one only needs to invert a low-dimensional matrix. On the contrary, if one does not employ a factor model, then $\mathrm{Var}(\mathbf{x}_m'\boldsymbol{\beta}_m^*|\mathbf{x}_{-m}) = \boldsymbol{\beta}_m^{*\prime}(\boldsymbol{\Sigma}_{x_m} - \boldsymbol{\Sigma}_{x_m,x_{-m}}\boldsymbol{\Sigma}_{x_{-m}}^{-1}\boldsymbol{\Sigma}_{x_{-m},x_m})\boldsymbol{\beta}_m^*$, where $\boldsymbol{\Sigma}_{x_m} = \mathrm{E}(\mathbf{x}_m^{\otimes 2})$, $\boldsymbol{\Sigma}_{x_m,x_{-m}} = \mathrm{E}(\mathbf{x}_m\mathbf{x}_{-m}')$, $\boldsymbol{\Sigma}_{x_{-m}} = \mathrm{E}(\mathbf{x}_{-m}^{\otimes 2})$, and $\boldsymbol{\Sigma}_{x_{-m},x_m} = \mathrm{E}(\mathbf{x}_{-m}\mathbf{x}_m')$. Consequently, a large dense matrix $\boldsymbol{\Sigma}_{x_{-m}}$ has to be inverted.

7 Numerical analysis

7.1 Test of a whole modality

We evaluate the empirical performance of the factor-adjusted score test of a whole modality proposed in Section 4. We generate $M = 3$ modalities, and consider two cases of $\mathbf{x}$. Specifically, for each modality $m = 1, 2, 3$, $\mathbf{x}_m$ are n i.i.d. random samples generated from $N_{p_m}(0, \boldsymbol{\Sigma}_m)$. For Case 1, $\boldsymbol{\Sigma}_m = \boldsymbol{\Lambda}_m\boldsymbol{\Lambda}_m' + 0.5\mathbf{I}_{p_m}$, where each column of $\boldsymbol{\Lambda}_m \in \mathcal{R}^{p_m \times K_m}$ is generated from $N_{p_m}(0, 2)$, and the number of factors $K_m = 2$. For Case 2, the diagonal elements of $\boldsymbol{\Sigma}_m$ equal 1 and the off-diagonal elements equal 0.4. Accordingly, in Case 1, $\mathbf{x}_m$ indeed follows a factor model setup, and in Case 2, although $\mathbf{x}_m$ does not strictly follow a factor model, its covariance matrix has spiked eigenvalues. In both cases, we aim to test if the first modality $\mathbf{x}_1$ is significantly associated with the response, i.e. $H_0: \beta_{11}^* = \cdots = \beta_{1p_1}^* = 0$. We then consider two types of alternatives. The first alternative is $H_{A1}: \beta_{11}^* = \cdots = \beta_{1p_1}^* = \delta/p$, where δ is a sequence approaching zero. As such, there is a weak signal in each variable of $\mathbf{x}_1$ and the overall signal is dense. The second alternative is $H_{A2}: \beta_{11}^* = \cdots = \beta_{15}^* = \delta/5, \beta_{16}^* = \cdots = \beta_{1p_1}^* = 0$. As such, the overall signal is sparse,

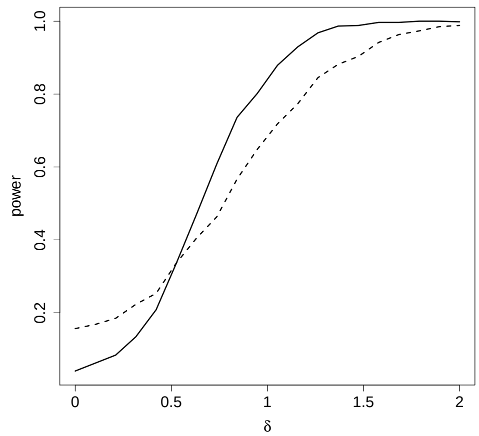


H_{A1}: $(n, p) = (100, 600)$ $\quad$ H_{A1}: $(n, p) = (200, 900)$

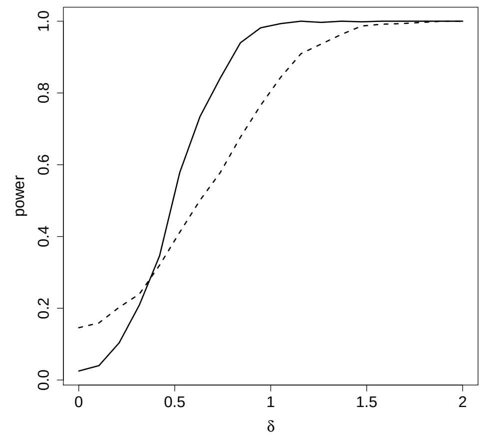
H_{A2}: $(n, p) = (100, 600)$ $\quad$ H_{A2}: $(n, p) = (200, 900)$

Figure 1: Empirical size and power of testing a whole modality for Case 1. The solid line is the proposed factor-adjusted score test, and the dashed line is the score test of Ning and Liu (2017).

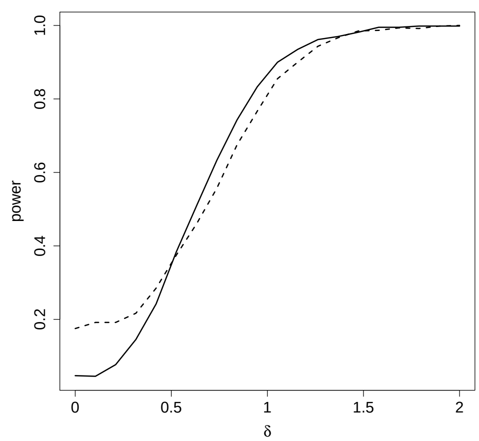
as all signals come only from the first 5 variables, whereas the rest do not associate with the response. For the other two modalities $\mathbf{x}_2$ and $\mathbf{x}_3$, we set $\beta^*_{21} = 1, \beta^*_{22} = 2, \beta^*_{23} = \cdots = \beta^*_{2p_2} = 0$, and $\beta^*_{31} = -1, \beta^*_{32} = -1, \beta^*_{33} = \cdots = \beta^*_{3p_3} = 0$. We generate the error ϵ as n i.i.d. samples from $N(0, 0.5)$, and generate y based on model (2). We set $p_1 = p_2 = p_3 = p/3$. We consider two combinations $(n, p) = (100, 600)$, and $(200, 900)$. We compare our test with the score test of Ning and Liu (2017), where the critical values are obtained by bootstrap.



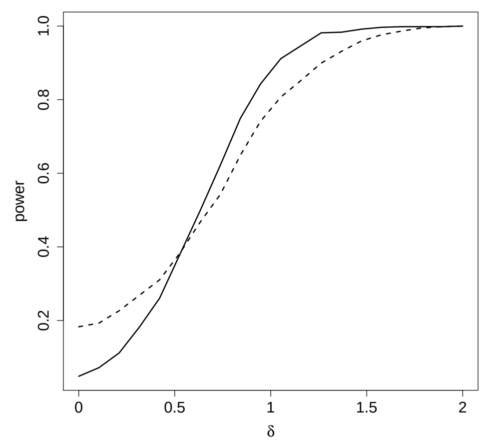
H_{A1}: $(n, p) = (100, 600)$



H_{A1}: $(n, p) = (200, 900)$



H_{A2}: $(n, p) = (100, 600)$



H_{A2}: $(n, p) = (200, 900)$

Figure 2: Empirical size and power of testing a whole modality for Case 2. The solid line is the proposed factor-adjusted score test, and the dashed line is the score test of Ning and Liu (2017).

We report the proportion of rejections of H_0 by both tests out of 600 data replications as we vary the value of δ. When $\delta = 0$, this gives the empirical size, and when $\delta > 0$, it gives the empirical power of the two tests. Figures 1 and 2 report the results for Cases 1 and 2, respectively. In both cases, we see that our proposed test controls the Type I error at the nominal level of 0.05 when $\delta = 0$. However, the test of Ning and Liu (2017) often yields an inflated size. This may be due to that their multiplier bootstrap method rejects the null hypothesis if the maximum of the decorrelated score functions of variables in that modality is greater than a threshold, and as such, it is easier to reject the null hypothesis. Moreover, our test achieves an as good or often a better power than the test of Ning and Liu (2017) as δ increases.

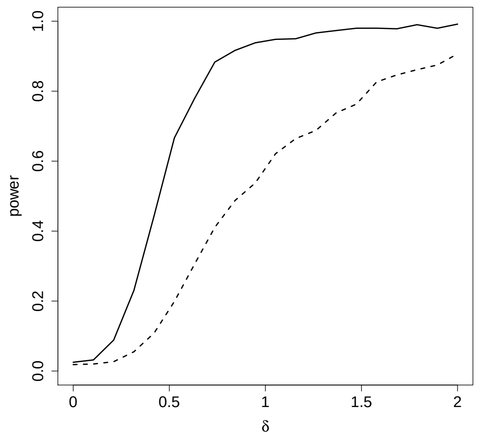
7.2 Test of a linear combination of predictors

We next evaluate the empirical performance of the factor-adjusted Wald test of a linear combination of predictors proposed in Section 5. We again generate $M = 3$ modalities and two cases of $\mathbf{x}$ similarly as in Section 7.1, except that in the second case we increase the off-diagonal elements of $\boldsymbol{\Sigma}_m$ to 0.8 for $m = 1, 2, 3$. We set $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*\prime}, \boldsymbol{\beta}_2^{*\prime}, \boldsymbol{\beta}_3^{*\prime})'$, $\boldsymbol{\beta}_1^* = (-2 + \delta, -1, 0, \ldots, 0)'$, $\boldsymbol{\beta}_2^* = (1 + \delta, 2, 0, \ldots, 0)'$, $\boldsymbol{\beta}_3^* = (1 + \delta, 1, 0, \ldots, 0)'$, and aim to test the linear combination of the first variable in each modality that $H_0 : \beta_{11}^* + \beta_{21}^* + \beta_{31}^* = 0$ versus $H_A : \beta_{11}^* + \beta_{21}^* + \beta_{31}^* \neq 0$. The rest of the simulation setup is the same as that in Section 7.1. We also compare our test with the partially penalized Wald test proposed in Shi et al. (2019).

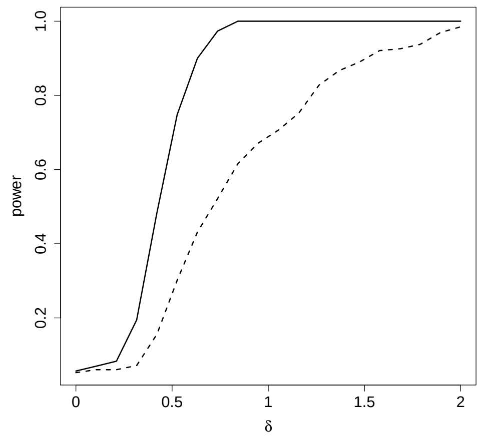
We again report the proportion of rejections of H_0 by both tests out of 600 data replications as we vary the value of δ. Figure 3 reports the results for both Cases 1 and 2. We see that, both tests achieve a good control of the Type I error at the nominal level when $\delta = 0$. Meanwhile, in both cases, our factor-adjusted Wald test achieves a much improved power as δ increases. The main reason is that, in this example, the variables are highly correlated with each other. The factor adjustment alleviates such high correlations, and yields better variable selection and estimation of the true regression coefficients, which in turn benefits the inference.

7.3 Multimodal neuroimaging analysis

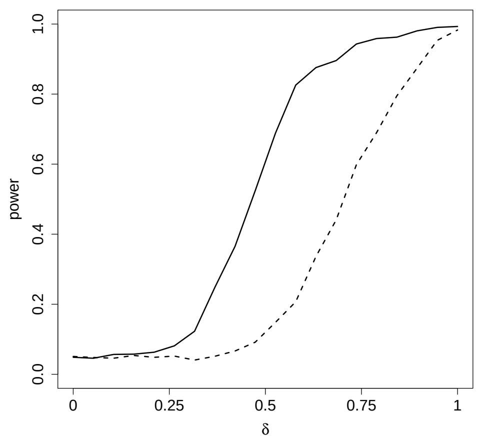
We illustrate our methods with a multimodal neuroimaging analysis to study Alzheimer's disease (AD). AD is an irreversible neurodegenerative disorder characterized by progressive impairment of cognitive and memory functions. It is the leading form of dementia in elderly subjects, and is the sixth leading cause of death in the United States. In 2018, AD affects over 5.5 million Americans, and without any effective treatment and prevention,



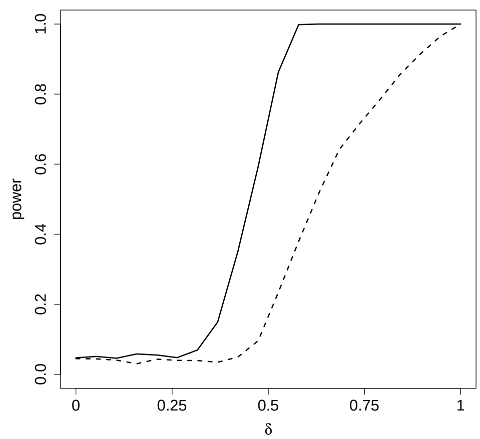
Case 1: $(n, p) = (100, 600)$



Case 1: $(n, p) = (200, 900)$



Case 2: $(n, p) = (100, 600)$



Case 2: $(n, p) = (200, 900)$

Figure 3: Empirical size and power of testing a linear combination of predictors. The solid line is the proposed factor-adjusted Wald test, and the dashed line is the partially penalized Wald test of Shi et al. (2019).

this number is projected to almost triple by 2050 (Alzheimer's Association, 2018). Tau is a hallmark pathological protein of AD, and is believed to be part of the driving mechanism of the disorder. It is present in the brains of both AD subjects and the elderly absent of dementia. Brain atrophy is another well known characteristic that differentiates between AD and normal aging. We studied a dataset with $n = 125$ subjects. Each subject received a positron emission tomography (PET) scan with AV-1451 tracer that measures accumulation

		R.rostantcing	L.superiorparietal	L.inferiorparietal	L.middletemporal
Coefficient	$\widehat{\boldsymbol{\beta}}_1$	0.07	0.12	0	0
	$\widehat{\boldsymbol{\beta}}_2$	0	0	-0.18	-0.01
p-value		0.05	0.03	0.001	0.09
		L.parahippocampal	L.rostantcing	R.parstriangularis	R.superiortemporal
Coefficient	$\widehat{\boldsymbol{\beta}}_1$	0	0	0	0
	$\widehat{\boldsymbol{\beta}}_2$	0.07	0.06	0.01	0.12
p-value		0.01	0.02	0.13	0.02
		R.supramarginal	R.temppole		
Coefficient	$\widehat{\boldsymbol{\beta}}_1$	0	0		
	$\widehat{\boldsymbol{\beta}}_2$	0.11	0.04		
p-value		0.02	0.04		

Table 1: The identified brain regions with the coefficient estimates and the corresponding p-values of the factor-adjusted Wald test for the significance of the brain regions.

of tau protein, as well as an anatomical magnetic resonance imaging (MRI) scan that measures brain grey matter cortical thickness. We mapped both types of images to a common brain atlas from Free Surfer, then summarized each PET and MRI image by a 58-dimensional vector, with each entry measuring the tau accumulation and cortical thickness of a particular brain region-of-interest (ROI), respectively. We removed some regions with quality issues for the PET images, which result in $p_1 = 51$ ROIs for PET, and $p_2 = 58$ ROIs for MRI, for each subject. In our integrative analysis, the tau and cortical thickness measurements form the two modalities $\mathbf{x}_1$ and $\mathbf{x}_2$. Memory score is a critical measure of cognitive decline for AD, and in our analysis, the memory score after removing potential age and sex effects is the response variable y.

We estimated the number of latent factors using the method of Bai and Li (2012), which concluded that there are $\widehat{K}_1 = 3$ factors in the tau modality and $\widehat{K}_2 = 1$ factor in the cortical thickness modality. We then estimated $\boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}^*$ using (4) with a SCAD penalty, where the tuning parameter was chosen by cross-validation. We then applied the three methods we develop in this article. We first tested the significance of the entire modality using the factor-adjusted score test in Section 4. The p-values are 8.5×10^{-7} and

5.9×10^{-3}, for testing the significance of tau and cortical thickness modality, respectively. As such, both modalities are clearly significantly associated with the memory outcome. We then report the estimated non-zero coefficients from our integrative factor model and their corresponding brain regions in Table 1. We further carried out the factor-adjusted Wald test in Section 5 to evaluate if the identified regions are significantly correlated with the outcome in either modality. We report the corresponding p-values in Table 1 as well. Our findings agree with the AD literature. For instance, the ROI with the smallest p-value we found is inferior parietal lobe, which is one of brain regions that is known to be associated with progression from healthy aging to AD (Greene and Killiany, 2010). Another significant ROI is parahippocampal gyrus, and cortical thinning of this region has been identified as an early biomarker of AD (Echávarri et al., 2011; Krumm et al., 2016). Finally, we evaluated the contribution of each individual modality. If we include the tau modality $\mathbf{x}_1$ in the model first, and add the cortical thickness modality $\mathbf{x}_2$ next, we have $\widehat{\sigma}_1^2 = \widehat{\mathrm{Var}}(\mathbf{x}_1' \boldsymbol{\beta}_1^*) = 0.11$, and $\widehat{\sigma}_{2|1}^2 = \widehat{\mathrm{Var}}(\mathbf{x}_2' \boldsymbol{\beta}_2^* | \mathbf{x}_1) = 0.18$. Correspondingly, $\widehat{\sigma}_1^2 / \widehat{\sigma}_y^2 = 14\%$, and $\widehat{\sigma}_{2|1}^2 / \widehat{\sigma}_y^2 = 24\%$. In other words, the tau modality explains 14% total variation in the response, and adding the cortical thickness modality explains an additional 24% total variation. On the other hand, if we include the cortical thickness modality $\mathbf{x}_2$ in the model first, and add the tau modality $\mathbf{x}_1$ next, we have $\widehat{\sigma}_2^2 = \widehat{\mathrm{Var}}(\mathbf{x}_2' \boldsymbol{\beta}_2^*) = 0.19$, and $\widehat{\sigma}_{1|2}^2 = \widehat{\mathrm{Var}}(\mathbf{x}_1' \boldsymbol{\beta}_1^* | \mathbf{x}_2) = 0.08$. Correspondingly, $\widehat{\sigma}_2^2 / \widehat{\sigma}_y^2 = 25\%$, and $\widehat{\sigma}_{1|2}^2 / \widehat{\sigma}_y^2 = 11\%$. In other words, the cortical thickness modality explains 25% total variation in the response, and adding the tau modality explains an additional 11% total variation.

References

Alzheimer's Association (2018). 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia* **14,** 367 – 429.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics* **40,** 436–465.

Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica* **70,** 191–221.

Bentkus, V. (2005). A Lyapunov-type bound in $\mathcal{R}^d$. *Theory of Probability & Its Applications* **49,** 311–323.

Cai, T. T., Ma, Z., and Wu, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41,** 3074–3110.

Echávarri, C., Aalten, P., et al. (2011). Atrophy in the parahippocampal gyrus as an early biomarker of Alzheimer's disease. *Brain Structure and Function* **215,** 265–271.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B.* **74,** 37–65.

Fan, J., Ke, Y., and Wang, K. (2016). Factor-adjusted regularized model selection. *arXiv:1612.08490* .

Fan, J., Liao, Y., Mincheva, M., and Jan, S. T. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B.* **75,** 603–680.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *Information Theory, IEEE Transactions* **57,** 5467–5484.

Gaynanova, I. and Li, G. (2019). Structural learning and integrative decomposition of multi-view data. *arXiv:1707.06573.*

Greene, S.J. and Killiany, R.J. (2011). Subregions of the inferior parietal lobule are affected in the progression to Alzheimer's disease. *Neurobiology of Aging* **31,** 1304–1311.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* **15,** 2869–2909.

Kapetanios, G. (2010). A testing procedure for determining the number of factors in approximate factor models with large datasets. *Journal of Business & Economic Statistics* **28,** 397–409.

Kneip, A. and Sarda, P. (2011). Factor models and variable selection in high-dimensional regression analysis. *The Annals of Statistics* **39,** 2410–2447.

Krumm, S., Kivisaari, S.L., et al. (2016). Cortical thinning of parahippocampal subregions in very early Alzheimer's disease. *Neurobiology of Aging* **38,** 188–196.

Li, G. and Gaynanova, I. (2018). A general framework for association analysis of heterogeneous data. *The Annals of Applied Statistics* **12,** 1700–1726.

Li, G. and Jung, S. (2017). Incorporating covariates into integrated factor analysis of multi-view data. *Biometrics* **73,** 1433–1442.

Li, G., Liu, X., and Chen, K. (2018). Integrative multi-view reduced-rank regression: Bridging group-sparse and low-rank models. *arXiv:1308.1479* .

Li, Q., Cheng, G., Fan, J., and Wang, Y. (2018). Embracing the blessing of dimensionality in factor models. *Journal of the American Statistical Association* **113,** 380–389.

Li, Y., Wu, F.-X., and Ngom, A. (2016). A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics* **19,** 325–340.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The Annals of Applied Statistics* **7,** 523–542.

Lock, E. F. and Li, G. (2018). Supervised multiway factorization. *Electronic Journal of Statistics* **12,** 1150.

Lynch, M. and Walsh, B. (1998). *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics* **41,** 772–801.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27,** 538–557.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* **45,** 158–195.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends® in Optimization* **1,** 127–239.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over L_q-balls. *IEEE Transactions on Information Theory* **57,** 6976–6994.

Richardson, S., Tseng, G. C., and Sun, W. (2016). Statistical methods in integrative genomics. *Annual Review of Statistics and Its Applications* **3,** 181–209.

Shen, R., Wang, S., and Mo, Q. (2013). Sparse integrative clustering of multiple omics data sets. *The Annals of Applied Statistics* **7,** 269–294.

Shi, C., Song, R., Chen, Z., and Li, R. (2019). Linear hypothesis testing for high dimensional generalized linear models. *The Annals of Statistics, to appear.*

Shu, H., Wang, X., and Zhu, H. (2019). D-CCA: A decomposition-based canonical correlation analysis for high-dimensional datasets. *Journal of the American Statistical Association*, to appear.

Sun, T. and Zhang, C.-H. (2013). Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research* **14,** 3385–3418.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B.* **58,** 267–288.

Uludag, K. and Roebroeck, A. (2014). General overview on the merits of multimodal neuroimaging data fusion. *Neuroimage* **102,** 3–10.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42,** 1166–1202.

Xue, F. and Qu, A. (2019). Integrating multi-source block-wise missing data in model selection. *arXiv:1901.03797* .

Yang, Z. and Michailidis, G. (2015). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* **32,** 1–8.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38,** 894–942.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B.* **76,** 217–242.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., and the Alzheimers Disease Neuroimaging Initiative (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* **55,** 856 – 867.

Zhang, Y., Tang, N., and Qu, A. (2019). Imputed factor regression for high-dimensional block-wise missing data. *Statistica Sinica*, to appear.