A GENERAL FRAMEWORK FOR ASSOCIATION ANALYSIS OF HETEROGENEOUS DATA

By Gen Li[1] and Irina Gaynanova

Columbia University and Texas A&M University

Multivariate association analysis is of primary interest in many applications. Despite the prevalence of high-dimensional and non-Gaussian data (such as count-valued or binary), most existing methods only apply to low-dimensional data with continuous measurements. Motivated by the Computer Audition Lab 500-song (CAL500) music annotation study, we develop a new framework for the association analysis of two sets of high-dimensional and heterogeneous (continuous/binary/count) data. We model heterogeneous random variables using exponential family distributions, and exploit a structured decomposition of the underlying natural parameter matrices to identify shared and individual patterns for two data sets. We also introduce a new measure of the strength of association, and a permutation-based procedure to test its significance. An alternating iteratively reweighted least squares algorithm is devised for model fitting, and several variants are developed to expedite computation and achieve variable selection. The application to the CAL500 data sheds light on the relationship between acoustic features and semantic annotations, and provides effective means for automatic music annotation and retrieval.

1. Introduction. With the advancement of measurement technologies, data acquisition becomes cheaper and easier. Often, data are collected from multiple sources or different platforms on the same set of samples, which are known as multi-view or multi-modal data. One of the main challenges associated with the analysis of multi-view data is that measurements from different sources may have heterogeneous types, such as continuous, binary, and count-valued. For instance, the motivating Computer Audition Lab 500-song (CAL500) data [Turnbull et al. (2007)] contain two sets of variables, acoustic features and semantic annotations, which are collected for 502 Western popular songs from the past 50 years. The acoustic features characterize the audio textures of a song, and are continuous variables obtained from well-developed signal processing methods [see, e.g., Logan (2000)]. The semantic annotations represent a song with a binary vector of labels over a multi-word vocabulary of semantic concepts. The labels correspond to different genres, usages, instruments, characteristics, and vocal types.

In large music databases, it is often desired to have computers automatically generate a short description for a novel song from its acoustic features (auto-tagging), or select relevant songs based on a multi-word semantic query (music retrieval) [Barrington et al. (2007), Bertin-Mahieux et al. (2008), Goto and Hirata (2004), Turnbull et al. (2007, 2008)]. The CAL500 study provides a well annotated music database to achieve these goals. The matched acoustic features and annotation profiles facilitate the investigation of the association between the two sets of variables.The association analysis may not only reveal how audio textures jointly affect listeners' subjective feelings, but also identify annotation patterns that can be used for music retrieval. As a result, it may give rise to new, effective auto-tagging and retrieval methods.

One of the most popular methods for the multivariate association analysis is the canonical correlation analysis (CCA) [Hotelling (1936)]. The CCA seeks linear combinations of the two sets of continuous variables with the maximal correlation. The loadings of the combinations offer insights into how the two sets of variables are related, whereas the resulting correlation is used to assess the strength of association. Furthermore, the canonical variables can be used for subsequent analyses such as regression [Luo et al. (2016)] and clustering [Chaudhuri et al. (2009)]. However, the standard CCA has many limitations. On the one hand, it implicitly assumes that both sets of variables are real-valued in order to make the linear combinations interpretable. Moreover, the Gaussian assumption is used to provide a probabilistic interpretation [Bach and Jordan (2005)]. That said, the CCA is not appropriate for non-Gaussian data, such as the binary annotations in the CAL500 study. On the other hand, the CCA suffers from overfitting for high-dimensional data. When the number of variables in either data set exceeds the sample size, the largest canonical correlation will always be one, resulting in misleading conclusions. Several extensions have been studied in the literature to address the overfitting issue, with sparsity regularization being the most common approach [Chen and Liu (2012), Chen et al. (2013), Witten, Tibshirani and Hastie (2009)]. These methods, however, are not directly applicable to non-Gaussian data.

To conduct the association analysis of the CAL500 data, we develop a new framework that accommodates high-dimensional heterogeneous variables. We call it the *Generalized Association Study* (GAS) framework. We model heterogeneous data types (binary/count/continuous) using exponential family distributions, and exploit a structured decomposition of the underlying natural parameter matrices to capture the dependency structure between the variables. The natural parameter matrices are specifically factorized into joint and individual structure, where the joint structure characterizes the association between the two data sets, and individual structure captures the remaining variation in each set. The proposed framework builds upon a low-rank model, which reduces the overfitting issue for high dimensional data. To our knowledge, this is the first attempt to generalize the multivariate association analysis to high dimensional non-Gaussian data from a frequentist perspective. We apply the method to the CAL500 data, and explicitly characterize the

dependency structure between the acoustic features and the semantic annotations. We further use the proposed framework to devise new procedures for auto-tagging and music retrieval. The resulting annotation performance is superior to existing methods.

The proposed model connects to the joint and individual variation explained (JIVE) model [Lock et al. (2013)] and the inter-battery factor analysis (IBFA) model [Browne (1979), Tucker (1958)] under the Gaussian assumption. Klami, Virtanen and Kaski (2010, 2013), Virtanen, Klami and Kaski (2011) extended the IBFA model to non-Gaussian data under the Bayesian framework and developed Bayesian CCA methods for the association analysis. However, the Bayesian methods require Gaussian priors for technical considerations, and are computationally prohibitive for large data. A major difference of the proposed method is that we treat the underlying natural parameters as fixed effects and exploit a frequentist approach to estimate them without imposing any prior distribution. The model parameters can be efficiently estimated using generalized linear models (GLM) and the algorithm scales well to large data. In addition, variable selection can be easily incorporated into the proposed framework to further facilitate interpretation. A similar idea has been explored in the context of mixed graphical models [Cheng et al. (2017), Landgraf and Lee (2015), Yang, Ning and Liu (2014)], which extend Gaussian graphical models to mixed data types. However, graphical models generally focus on characterizing relations between variables rather than data sets, and thus are not directly suitable for the purpose of music annotation and retrieval.

Another unique contribution of the paper is that we introduce a new measure of the strength of association between the two heterogeneous data sets: the *association coefficient*. We devise a permutation-based test which formally assesses the significance of association and provides a *p*-value. We apply the methods to the CAL500 data, and identify a statistically significant, yet moderate, association between the acoustic features and the semantic annotations. The statistical significance warrants the analysis of the dependency structure between the heterogeneous data types. The moderate association may partially explain why auto-tagging and query-by-semantic-description are challenging problems, and no existing machine learning method provides extraordinary performance [Bertin-Mahieux et al. (2008), Turnbull et al. (2008)].

The rest of the paper is organized as follows. In Section 2, we introduce the model and discuss identifiability conditions under the GAS framework. In Section 3, we describe the new association coefficient and a permutation-based hypothesis test for the significance of association. In Section 4, we elaborate the model fitting procedure. In Section 5, we apply the proposed framework to the CAL500 data, and discuss new procedures for auto-tagging and music retrieval. In Section 6, we conduct comprehensive simulation studies to compare our approach with existing methods. Discussion and concluding remarks are provided in Section 7. Proofs, technical details of the algorithm, a detailed description of the rank estimation procedure, and additional simulation results can be found in the Supplementary Material [Li and Gaynanova (2018)].

2. Generalized association study framework. In this section, we first introduce a statistical model for characterizing the dependency structure between two non-Gaussian data sets. Then we discuss the identifiability of the proposed model.

2.1. *Model.* Let X_1 and X_2 be two data matrices of size $n \times p_1$ and $n \times p_2$, respectively, with rows being the samples (matched between the matrices) and columns being the variables. We assume the entries of each data matrix are realizations of univariate random variables from a single-parameter exponential family distribution (e.g., Gaussian, Poisson, Bernoulli). In particular, the random variables may follow different distributions in different matrices. The probability density function of each random variable x takes the form

$$f(x|\theta) = h(x) \exp\{x\theta - b(\theta)\},$$

where $\theta \in \mathbb{R}$ is a natural parameter, $b(\cdot)$ is a convex cumulant function, and $h(\cdot)$ is a normalization function. The expectation of the random variable is $\mu = b'(\theta)$. Following the notation in the GLM framework, the canonical link function is defined as $g(\mu) = b'^{-1}(\mu)$. The notation for some commonly used exponential family distributions is given in Table 1.

Each random variable in the data matrix X_k corresponds to a unique underlying natural parameter, and all the natural parameters form an $n \times p_k$ parameter matrix $\mathbf{\Theta}_k \in \mathbb{R}^{n \times p_k}$. The univariate random variables are assumed conditionally independent, given the underlying natural parameters. The relation among the random variables is captured by the intrinsic patterns of the natural parameter matrices $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$, which serve as the building block of the proposed model. We remark that the conditional independence assumption given underlying natural parameters is commonly used in the literature for modeling multivariate non-Gaussian data. See, Goldsmith, Zipunnikov and Schrack (2015), Landgraf and Lee (2015), She (2013), Zoh et al. (2016), for example. On the one hand, univariate exponential family distributions are more tractable than the multivariate counterparts [Johnson, Kotz and Balakrishnan (1997)]. Other than the multivariate Gaussian distribution, multivariate exponential family distributions are generally less studied and hard to use. On the other hand, the entry-wise natural parameters can be used to capture

TABLE 1
The notation for some commonly used exponential family distributions

	Mean μ	Natural parameter θ	$b(\theta)$	$g(\mu)$
Gaussian (with unit variance)	μ	μ	$\frac{\theta^2}{2}$	μ
Poisson	λ	$\log \lambda$	$\exp(\theta)$	$\log(\mu)$
Bernoulli	p	$\log \frac{p}{1-p}$	$\log\{1 + \exp(\theta)\}$	$\log \frac{\mu}{1-\mu}$

the statistical dependency in multivariate settings, acting similarly to a covariance matrix. For example, Collins, Dasgupta and Schapire (2001) provided an alternative interpretation of the principal component analysis (PCA) using the low rank approximation to the natural parameter matrix.

Under the independence assumption, each entry of X_k follows an exponential family distribution with the probability density function $f_k(\cdot)$ and the corresponding natural parameter matrix Θ_k. To characterize the joint structure between the two data sources and the individual structure within each data source, we model Θ_1 and Θ_2 as

$$(2.1) \qquad \begin{cases} \Theta_1 = \mathbf{1}\mu_1^T + U_0 V_1^T + U_1 A_1^T, \\ \Theta_2 = \mathbf{1}\mu_2^T + U_0 V_2^T + U_2 A_2^T. \end{cases}$$

Each parameter matrix is decomposed into three parts: the intercept (the first term), the joint structure (the second term), and the individual structure (the third term). In particular, $\mathbf{1}$ is an length-n vector of all ones and μ_k is a length-p_k *intercept* vector for Θ_k. Let r_0 and r_k denote the joint and individual ranks respectively, where $r_0 \leq \min(n, p_1, p_2)$ and $r_k \leq \min(n, p_k)$. Then, U_0 is an $n \times r_0$ shared *score* matrix between the two parameter matrices; $(V_1^T, V_2^T)^T$ is a $(p_1 + p_2) \times r_0$ shared *loading* matrix, where V_k corresponds to Θ_k only; U_k and A_k are $n \times r_k$ and $p_k \times r_k$ individual score and loading matrices for Θ_k, respectively.

The decomposition of the natural parameter matrices in (2.1) has an equivalent form from the matrix factorization perspective. More specifically,

$$(\Theta_1, \Theta_2) = (\mathbf{1}, U_0, U_1, U_2) \begin{pmatrix} \mu_1^T & \mu_2^T \\ V_1^T & V_2^T \\ A_1^T & \mathbf{0} \\ \mathbf{0} & A_2^T \end{pmatrix},$$

where $\mathbf{0}$ represents any zero matrix of compatible size. This structured decomposition sheds light on the association and specificity of the two data sources. In particular, the joint score U_0 can be viewed as common latent factors that capture the association between the two data sets. The joint loading for each data set characterizes how the latent factors are constructed using variables in that data set. In practice, when it is reasonable to assume the existence of some shared factors underlying two sets of variables, the proposed model is suitable for modeling their association. For instance, in our motivating music annotation example, it is logical to believe there are latent factors characterizing the emotional intensity (e.g., from calming to arousing) and attitude spectrum (e.g., from cheerful to hateful) of songs in both the audio feature data and the annotation profile data. Thus we can use Model (2.1) to identify the common latent factors, and further infer association.

Loosely speaking, if the joint structure dominates the decomposition, the two sets of variables are deemed highly associated. On the contrary, if the individual structure is dominant, they are less connected. A more rigorous measure of the association is given in Section 3.

2.2. *Connection to existing models.* Under the Gaussian assumption on X_1 and X_2, Model (2.1) is identical to the JIVE model with two data sets [Lock et al. (2013)]:

$$X_1 = \mathbf{1}\mu_1^T + U_0 V_1^T + U_1 A_1^T + E_1,$$
$$X_2 = \mathbf{1}\mu_2^T + U_0 V_2^T + U_2 A_2^T + E_2,$$

where E_1 and E_2 are additive noise matrices. JIVE is an example of linked component models [Zhou et al. (2016b)], where the dependency between two data sets is characterized by the presence of fixed shared latent components (i.e., U_0). When the shared components are absent, JIVE reduces to individual PCA models for X_1 and X_2. When the individual components are absent, JIVE reduces to a consensus PCA model [Westerhuis, Kourti and MacGregor (1998)]. These models are closely related to the factor analysis, and the main difference is the deterministic (rather than probabilistic) treatment of latent components. If we substitute the fixed parameters U_0 and U_k with Gaussian random variables, Model (2.1) coincides with the IBFA model [Browne (1979), Tucker (1958)]. The deterministic approach, however, allows us to interpret JIVE as a multi-view generalization of the standard PCA. While explicitly designed for modeling associations between two data sets, CCA cannot take into account individual latent components. As a result, it has been shown that linked component models often outperform CCA in the estimation of joint associations [Jia, Salzmann and Darrell (2010), Trygg and Wold (2003), Zhou et al. (2016a)]. For further comparison between CCA and JIVE, we refer the reader to [Lock et al. (2013)].

The proposed framework extends linked component models to the exponential family distributions. Rewriting Model (2.1) with respect to each entry of X_1 and X_2 (denoted by x_{1ij} and x_{2ik}) leads to

$$x_{1ij} \sim f_1(\theta_{1ij}), \qquad x_{2ik} \sim f_2(\theta_{2ik})$$

with

$$\theta_{1ij} = \mu_{1j} + \sum_{r=1}^{r_0} u_{0ir} v_{1jr} + \sum_{l=1}^{r_1} u_{1il} a_{1jl},$$

$$\theta_{2ik} = \mu_{2j} + \sum_{r=1}^{r_0} u_{0ir} v_{2kr} + \sum_{m=1}^{r_2} u_{2im} a_{2km},$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are exponential family probability density functions associated with X_1 and X_2; and $u_{0ir}, u_{1il}, u_{2im}, v_{1jr}, v_{2kr}, a_{1jl}, a_{2km}$ are elements

of U_0, U_1, U_2, V_1, V_2, A_1, and A_2, respectively. The above display reveals that U_0, U_1, U_2 can be viewed as fixed latent factors with U_0 being shared across both data sets, and U_1, U_2 being data set-specific. As such, this model is closely connected to the factor analysis in the context of generalized linear models. The factors are used to model the means of random variables through the canonical link functions rather than directly. The deterministic treatment allows us to interpret our model as a multi-view generalization of the exponential PCA [Collins, Dasgupta and Schapire (2001)], similar to JIVE as a multi-view generalization of the standard PCA.

2.3. *Identifiability.* To ensure the identifiability of Model (2.1), we consider the following regularity conditions:

- The columns of the individual score matrices (U_1 and U_2) are linearly independent; the intercept (μ_k) and the columns of the joint and individual loading matrices (V_k and A_k) corresponding to each data type are linearly independent;
- The score matrices are column-centered [i.e., $\mathbf{1}^T(U_0, U_1, U_2) = \mathbf{0}$], and the column space of the joint score matrix is orthogonal to that of the individual score matrices [i.e., $U_0^T(U_1, U_2) = \mathbf{0}$];
- Each score matrix has orthogonal columns, and each loading matrix has orthonormal columns (i.e., $V_1^T V_1 + V_2^T V_2 = I$, $A_1^T A_1 = I$ and $A_2^T A_2 = I$, where I is an identity matrix of compatible size).

The first condition ensures that the joint and individual ranks are correctly specified. The second condition orthogonalizes the intercept, the joint and the individual patterns. The last condition rules out the arbitrary rotation and rescaling of each decomposition, if the column norms of respective score matrices are distinct (this is almost always true in practice). We remark that the orthonormality condition for the concatenated joint loadings in $(V_1^T, V_2^T)^T$ is more general than separate orthonormality conditions for V_1 and V_2, and is beneficial for modeling data with different scales and structures. Under the above conditions, Model (2.1) is uniquely defined up to trivial column reordering and sign switches. The rigorous proof of the model identifiability partially attributes to the Theorem 1.1 in the Supplementary Material of Lock et al. (2013). For completeness, we restate the theorem under our framework:

PROPOSITION 2.1. *Let*

$$\begin{cases} \Theta_1 = J_1 + B_1, \\ \Theta_2 = J_2 + B_2, \end{cases}$$

$J = (J_1, J_2)$ *and* $B = (B_1, B_2)$, *where* $\operatorname{rank}(J) = r_0$ *and* $\operatorname{rank}(B_k) = r_k$ *for* $k = 1, 2$. *Suppose the model ranks are correctly specified, that is,* $\operatorname{rank}(B) = r_1 + r_2$ *and* $\operatorname{rank}(\Theta_k) = r_0 + r_k$ *for* $k = 1, 2$. *There exists a unique parameter set* $\{J_1, J_2, B_1, B_2\}$ *satisfying* $J^T B = \mathbf{0}$.

In Model (2.1), we have $\boldsymbol{J}_k = \mathbf{1}\boldsymbol{\mu}_k^T + \boldsymbol{U}_0\boldsymbol{V}_k^T$ and $\boldsymbol{B}_k = \boldsymbol{U}_k\boldsymbol{A}_k^T$ ($k = 1, 2$). Our first identifiability condition is equivalent to the rank prerequisite in Proposition 2.1. The second condition guarantees $\boldsymbol{J}^T\boldsymbol{B} = \mathbf{0}$. Hence the joint and individual patterns of our model are uniquely defined. Furthermore, our last identifiability condition is the standard condition that guarantees the uniqueness of the singular value decomposition (SVD) of a matrix [Golub and Van Loan (2013)].

3. Association coefficient and permutation test.

3.1. *Association coefficient.* Model (2.1) specifies the joint and individual structure of the natural parameter matrices underlying the two data sets. The relative weights of the joint structure can be used to measure the strength of association between the two data sources. Intuitively, if the joint structure dominates the individual structure, the latent generating schemes of the two data sets are coherent. Consequently, the two data sources are deemed highly associated. On the contrary, if the joint signal is weak, each data set roughly follows an independent EPCA generative model [Collins, Dasgupta and Schapire (2001)], and hence the two data sources are unrelated. To formalize this idea, we define an association coefficient between the two data sets as follows.

DEFINITION 3.1. Let $\boldsymbol{X}_1 \in \mathbb{R}^{n \times p_1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{n \times p_2}$ be two data sets with n matched samples, and assume $\boldsymbol{X}_k$ ($k = 1, 2$) follows an exponential family distribution with the entrywise underlying natural parameter matrix $\boldsymbol{\Theta}_k$. Let $\overline{\boldsymbol{\Theta}}_k$ be the column centered $\boldsymbol{\Theta}_k$. The *association coefficient* between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is defined as

$$(3.1) \qquad \rho(\boldsymbol{X}_1, \boldsymbol{X}_2) = \frac{\|\overline{\boldsymbol{\Theta}}_1^T \overline{\boldsymbol{\Theta}}_2\|_\star}{\|\overline{\boldsymbol{\Theta}}_1\|_\mathbb{F}\|\overline{\boldsymbol{\Theta}}_2\|_\mathbb{F}},$$

where $\|\cdot\|_\star$ and $\|\cdot\|_\mathbb{F}$ represent the nuclear norm and Frobenius norm of a matrix, respectively. In particular, under Model (2.1) with the identifiability conditions, the association coefficient has the expression

$$\rho(\boldsymbol{X}_1, \boldsymbol{X}_2) = \frac{\|\boldsymbol{V}_1\boldsymbol{U}_0^T\boldsymbol{U}_0\boldsymbol{V}_2^T + \boldsymbol{A}_1\boldsymbol{U}_1^T\boldsymbol{U}_2\boldsymbol{A}_2^T\|_\star}{\|\boldsymbol{U}_0\boldsymbol{V}_1^T + \boldsymbol{U}_1\boldsymbol{A}_1^T\|_\mathbb{F}\|\boldsymbol{U}_0\boldsymbol{V}_2^T + \boldsymbol{U}_2\boldsymbol{A}_2^T\|_\mathbb{F}}.$$

The definition of the association coefficient (3.1) only depends on the natural parameter matrix underlying each data set. It does not rely on our model assumption. Thus it is applicable in a broad context. Furthermore, the association coefficient satisfies the following properties. The proof can be found in Section A of the Supplementary Material [Li and Gaynanova (2018)].

PROPOSITION 3.2. (i) *The association coefficient $\rho(\boldsymbol{X}_1, \boldsymbol{X}_2)$ is bounded between* 0 *and* 1.

(ii) $\rho(X_1, X_2) = 0$ *if and only if the column spaces of* $\overline{\Theta}_1$ *and* $\overline{\Theta}_2$ *are mutually orthogonal.*

(iii) $\rho(X_1, X_2) = 1$ *if* $\overline{\Theta}_1$ *and* $\overline{\Theta}_2$ *have the same left singular vectors and proportional singular values.*

The first property puts the association coefficient on scale, making it similar to the conventional notion of correlation. A smaller value means weaker association, and vice versa. The second and third properties establish the conditions for "no association" and "perfect association," respectively. We remark that the second property provides a necessary and sufficient condition for $\rho(X_1, X_2) = 0$, while the third property only provides a sufficient condition for $\rho(X_1, X_2) = 1$. In the context of Model (2.1), we have the following corollary.

COROLLARY 3.3. *Suppose Model* (2.1) *has correctly specified ranks and satisfies the identifiability conditions. Then:*

(i) $\rho(X_1, X_2) = 0$, *if and only if* $U_0 = 0$ *and* $U_1^T U_2 = 0$;
(ii) $\rho(X_1, X_2) = 1$, *if* $U_1 = 0, U_2 = 0, V_1^T V_1 = cI$ *and* $V_2^T V_2 = (1-c)I$ *for some constant* $0 < c < 1$.

Conceptually, the association coefficient is zero when the joint structure is void and the individual patterns are mutually orthogonal in both data sets. Perhaps less obvious are the conditions for the two data sets to have the association coefficient exactly equal to one. Not only the individual structure does not exist, but the columns of V_1 (and V_2) must be mutually orthogonal with the same norm. It turns out the additional rigor is necessary. It reduces the risk of overestimating the association under model misspecification. See Section A of the Supplementary Material [Li and Gaynanova (2018)] for some concrete examples.

3.2. *Permutation test.* To formally assess the statistical significance of the association between X_1 and X_2, we consider the following hypothesis test:

$$H_0 : \rho(X_1, X_2) = 0 \quad \text{vs} \quad H_1 : \rho(X_1, X_2) > 0.$$

We use the sample version of the association coefficient $\rho(X_1, X_2)$ as the test statistic, and exploit a permutation-based testing procedure.

More specifically, assume $\overline{\Theta}_1$ and $\overline{\Theta}_2$ are estimated from data (see Section 4 for parameter estimation). The original test statistic, denoted by ρ_0, can be obtained from (3.1). Now we describe the permutation procedure. Let P_π be an $n \times n$ permutation matrix with the random permutation $\pi : \{1, \ldots, n\} \mapsto \{1, \ldots, n\}$. We keep X_1 fixed and permute the rows of X_2 based on π. As a result, the association between the two data sets is removed while the respective structure is reserved. The corresponding association coefficient for the permuted data, denoted by ρ_π, is a random sample under the null hypothesis. Because the natural parameters are

defined individually and permuted along with X_2, the column centered natural parameter matrix for $P_\pi X_2$ is $P_\pi \overline{\Theta}_2$. Thus, we directly obtain the expression of ρ_π as

$$\rho_\pi = \frac{\|\overline{\Theta}_1^T P_\pi \overline{\Theta}_2\|_\star}{\|\overline{\Theta}_1\|_\mathbb{F} \|P_\pi \overline{\Theta}_2\|_\mathbb{F}} = \frac{\|\overline{\Theta}_1^T P_\pi \overline{\Theta}_2\|_\star}{\|\overline{\Theta}_1\|_\mathbb{F} \|\overline{\Theta}_2\|_\mathbb{F}}.$$

We repeat the permutation procedure multiple times and get a sampling distribution of the association coefficient under the null. Consequently, the empirical p-value is calculated as the proportion of permuted values greater than or equal to the original test statistic ρ_0. A small p-value warrants further investigation on the dependency structure between the two data sets. More details of the operational characteristics of the permutation test can be found in Section I of the Supplementary Material [Li and Gaynanova (2018)].

4. Model fitting algorithm. In this section, we elaborate an alternating algorithm to estimate the parameters in Model (2.1). We show that the model fitting procedure can be formulated as a collection of GLM fitting problems. We also discuss how to incorporate variable selection into the framework via a regularization approach. When fitting the model, we assume the joint and individual ranks are fixed. We briefly introduce how to select the ranks at the end of this section. A more detailed data-driven rank selection approach is presented in Section D of the Supplementary Material [Li and Gaynanova (2018)].

4.1. *Alternating iteratively reweighted least square.* The model parameters in (2.1) consist of the intercept μ_k, the joint score U_0, the individual score U_k, the joint loading V_k, and the individual loading A_k ($k = 1, 2$). To estimate the parameters, we maximize the joint log likelihood of the observed data X_1 and X_2, denoted by $\ell(X_1, X_2 | \Theta_1, \Theta_2)$. Under the independence assumption, the joint log likelihood can be written as the summation of the individual log likelihoods for each value. Namely, we have

$$(4.1) \qquad \ell(X_1, X_2 | \Theta_1, \Theta_2) = \sum_{i=1}^{n} \sum_{j=1}^{p_1} \ell_1(x_{1,ij} | \theta_{1,ij}) + \sum_{i=1}^{n} \sum_{j=1}^{p_2} \ell_2(x_{2,ij} | \theta_{2,ij}),$$

where $X_k = (x_{k,ij})$ and $\Theta_k = (\theta_{k,ij})$, and ℓ_k is the log likelihood function for the kth distribution ($k = 1, 2$). In particular, Θ_1 and Θ_2 have the structured decomposition in (2.1). We estimate the parameters in a block-wise coordinate descent fashion: we alternate the estimation between the joint and the individual structure, and between the scores and the loadings (with the intercepts), until convergence.

More specifically, we first fix the joint structure $\{U_0, V_1, V_2\}$, and estimate the individual structure for each data set. Since the first term in (4.1) only involves $\{\mu_1, U_1, A_1\}$, and the second term only involves $\{\mu_2, U_2, A_2\}$, the parameter estimation is separable. We focus on the first term, and the second term can be updated

similarly. We first fix $\boldsymbol{\mu}_1$ and $\boldsymbol{A}_1$ to estimate $\boldsymbol{U}_1$. Let $\boldsymbol{u}_{k,(i)}$ be the column vector of the ith row of $\boldsymbol{U}_k$ ($k = 0, 1, 2$). The column vector of the ith row of $\boldsymbol{\Theta}_1$, denoted by $\boldsymbol{\theta}_{1,(i)}$, can be expressed as

$$\boldsymbol{\theta}_{1,(i)} = \boldsymbol{\mu}_1 + \boldsymbol{V}_1 \boldsymbol{u}_{0,(i)} + \boldsymbol{A}_1 \boldsymbol{u}_{1,(i)},$$

where everything is fixed except for $\boldsymbol{u}_{1,(i)}$. Noticing that the ith row of $\boldsymbol{X}_1$ (i.e., $\boldsymbol{x}_{1,(i)}$) and $\boldsymbol{\theta}_{1,(i)}$ satisfy

$$\mathbb{E}(\boldsymbol{x}_{1,(i)}) = b'_1(\boldsymbol{\theta}_{1,(i)}),$$

we exactly obtain a GLM with the canonical link. Namely, $\boldsymbol{x}_{1,(i)}$ is a generalized response vector; $\boldsymbol{A}_1$ is a $p_1 \times r_1$ predictor matrix; $\boldsymbol{\mu}_1 + \boldsymbol{V}_1 \boldsymbol{u}_{0,(i)}$ is an offset; $\boldsymbol{u}_{1,(i)}$ is a coefficient vector. The estimate of $\boldsymbol{u}_{1,(i)}$ can be obtained via an iteratively reweighted least squares (IRLS) algorithm [McCullagh and Nelder (1989)]. Furthermore, different rows of $\boldsymbol{U}_1$ can be estimated in parallel. Overall, the estimation of $\boldsymbol{U}_1$ is formulated as n parallel GLM fitting problems. Once $\boldsymbol{U}_1$ is estimated, we fix $\boldsymbol{U}_1$ and formulate the estimation of $\boldsymbol{\mu}_1$ and $\boldsymbol{A}_1$ as p_1 GLMs in a similar fashion. Consequently, we update the estimate of the individual structure.

Now we estimate the joint structure with fixed individual structure. When the joint score $\boldsymbol{U}_0$ is fixed, the estimation of $\{\boldsymbol{\mu}_1, \boldsymbol{V}_1\}$ and $\{\boldsymbol{\mu}_2, \boldsymbol{V}_2\}$ resembles the estimation of the individual counterparts. With fixed $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{V}_1, \boldsymbol{V}_2\}$, the estimation of $\boldsymbol{U}_0$ is slightly different, because it is shared by two data types with different distributions. Let $\boldsymbol{\theta}_{0,(i)} = (\boldsymbol{\theta}_{1,(i)}^T, \boldsymbol{\theta}_{2,(i)}^T)^T$ be a column vector concatenating the column vectors of the ith rows of $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$. Then we have

$$\boldsymbol{\theta}_{0,(i)} = (\boldsymbol{\mu}_1^T + \boldsymbol{u}_{1,(i)}^T \boldsymbol{A}_1^T, \boldsymbol{\mu}_2^T + \boldsymbol{u}_{2,(i)}^T \boldsymbol{A}_2^T)^T + \boldsymbol{V}_0 \boldsymbol{u}_{0,(i)},$$

where $\boldsymbol{V}_0 = (\boldsymbol{V}_1^T, \boldsymbol{V}_2^T)^T$ is the concatenated joint loading matrix. Notice that

$$\mathbb{E}(\boldsymbol{x}_{1,(i)}) = b'_1(\boldsymbol{\theta}_{1,(i)}), \qquad \mathbb{E}(\boldsymbol{x}_{2,(i)}) = b'_2(\boldsymbol{\theta}_{2,(i)}).$$

The formula corresponds to a nonstandard GLM where the response consists of observations from different distributions, and different link functions are used correspondingly. Following the standard GLM model fitting algorithm verbatim, we obtain a slightly modified version of the IRLS algorithm to address this problem. More details can be found in Section B of the Supplementary Material [Li and Gaynanova (2018)].

The separately estimated parameters, denoted by $\{\widehat{\boldsymbol{\mu}_1}, \widehat{\boldsymbol{\mu}_2}, \widehat{\boldsymbol{U}_0}, \widehat{\boldsymbol{U}_1}, \widehat{\boldsymbol{U}_2}, \widehat{\boldsymbol{V}_1}, \widehat{\boldsymbol{V}_2}, \widehat{\boldsymbol{A}_1}, \widehat{\boldsymbol{A}_2}\}$, may not satisfy the identifiability conditions in Section 2.3. In order to find an equivalent set of parameters satisfying the conditions, we conduct the following normalization procedure after each iteration. We first project the columns of the individual scores $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ to the orthogonal complement of the column space of $(\mathbf{1}, \boldsymbol{U}_0)$. The obtained individual score matrices are denoted by $\boldsymbol{U}_1^\star$ and $\boldsymbol{U}_2^\star$, which are column centered and orthogonal to the columns in $\boldsymbol{U}_0$. The new

individual patterns are $U_1^\star \widehat{A_1}^T$ and $U_2^\star \widehat{A_2}^T$ accordingly. To rule out arbitrary rotations and scale changes, we apply the SVD to each individual structure, and let the left singular vectors to absorb the singular values. As a result, we have

$$\widetilde{U}_1 \widetilde{A}_1^T = U_1^\star \widehat{A_1}^T, \qquad \widetilde{U}_2 \widetilde{A}_2^T = U_2^\star \widehat{A_2}^T,$$

where $\{\widetilde{U}_1, \widetilde{U}_2, \widetilde{A}_1, \widetilde{A}_2\}$ satisfies the identifiability conditions. Next, we add the remaining individual structure to the joint structure, and obtain the new joint structure as

$$\left(\mathbf{1}\widehat{\mu_1}^T + \widehat{U}_0\widehat{V}_1^T + \widehat{U}_1\widehat{A}_1^T - \widetilde{U}_1\widetilde{A}_1^T, \mathbf{1}\widehat{\mu_2}^T + \widehat{U}_0\widehat{V}_2^T + \widehat{U}_2\widehat{A}_2^T - \widetilde{U}_2\widetilde{A}_2^T\right).$$

Denote the new column mean vector as $(\widetilde{\mu}_1^T, \widetilde{\mu}_2^T)^T$, and center each column of the above joint structure. Subsequently, we apply SVD to the column-centered joint structure and obtain the new joint score $\widetilde{U}_0$ and joint loading $(\widetilde{V}_1^T, \widetilde{V}_2^T)^T$. As a result, the new parameter set $\{\widetilde{\mu}_1, \widetilde{\mu}_2, \widetilde{U}_0, \widetilde{U}_1, \widetilde{U}_2, \widetilde{V}_1, \widetilde{V}_2, \widetilde{A}_1, \widetilde{A}_2\}$ satisfies all the conditions, and provides the same likelihood value as the original parameter set.

In summary, we devise an alternating algorithm to estimate the model parameters. Each iteration is formulated as a set of GLMs, fitted by the IRLS algorithm. A step-by-step summary is provided in Algorithm 1. Because the likelihood value in (4.1) is nondecreasing in each optimization step, and remains constant in the normalization step, the algorithm is guaranteed to converge. More formally, we have the following proposition.

PROPOSITION 4.1. *In each iteration of Algorithm* 1, *the log likelihood* (4.1) *is monotonically nondecreasing. If the likelihood function is bounded, the estimates always converge to some stationary point* (*including infinity*).

Since the overall algorithm is iterative, we further substitute the IRLS algorithm with a one-step approximation with warm start to enhance computational efficiency. A detailed description is provided in Section C of the Supplementary Material [Li and Gaynanova (2018)]. In our numerical studies, we observe that the one-step approximation algorithm almost always converges to the same values as the full algorithm, but is several fold faster (see Section 6).

4.2. *Variable selection.* In practice, it is often desirable to incorporate variable selection into parameter estimation to facilitate interpretation, which is especially relevant when the number of variables is high. Various regularization frameworks and sparsity methods have been extensively studied in the literature. See Hastie, Tibshirani and Wainwright (2015) and references therein.

Since Model (2.1) is primarily used to investigate the association between the two data sets, it is of great interest to perform variable selection when estimating

---

Algorithm 1 The Alternating IRLS Algorithm for Fitting Model (2.1)

Initialize $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{U}_0, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{A}_1, \boldsymbol{A}_2\}$;
while The likelihood (4.1) has not reached convergence **do**

- Fix the joint structure $\{\boldsymbol{U}_0, \boldsymbol{V}_1, \boldsymbol{V}_2\}$
    - Fix $\{\boldsymbol{\mu}_1, \boldsymbol{A}_1\}$, and estimate each row of $\boldsymbol{U}_1$ via parallel GLM
    - Fix $\boldsymbol{U}_1$, and estimate each row of $(\boldsymbol{\mu}_1, \boldsymbol{A}_1)$ via parallel GLM
    - Fix $\{\boldsymbol{\mu}_2, \boldsymbol{A}_2\}$, and estimate each row of $\boldsymbol{U}_2$ via parallel GLM
    - Fix $\boldsymbol{U}_2$, and estimate each row of $(\boldsymbol{\mu}_2, \boldsymbol{A}_2)$ via parallel GLM
- Fix the individual structure $\{\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{A}_1, \boldsymbol{A}_2\}$
    - Fix $\boldsymbol{U}_0$, and estimate each row of $(\boldsymbol{\mu}_1, \boldsymbol{V}_1)$ via parallel GLM
    - Fix $\boldsymbol{U}_0$, and estimate each row of $(\boldsymbol{\mu}_2, \boldsymbol{V}_2)$ via parallel GLM
    - Fix $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{V}_1, \boldsymbol{V}_2\}$, and estimate each row of $\boldsymbol{U}_0$ via a modified IRLS algorithm in parallel
- Normalize the estimated parameters to retrieve the identifiability conditions

end while

---

the joint structure. In particular, sparse $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ facilitate model interpretability. The variables corresponding to nonzero joint loading entries can be used to interpret the association between the two data sources.

In order to achieve variable selection in the estimation, we modify the normalization step in each iteration of the model fitting algorithm. In particular, we substitute the SVD of the centered joint structure with the FIT-SSVD method developed by Yang, Ma and Buja (2014). The FIT-SSVD method provides sparse estimation of the singular vectors via soft or hard thresholding, while maintaining the orthogonality among the vectors. By default, an asymptotic threshold is used to automatically determine the sparsity level for each data set. Consequently, the method is directly embedded into our algorithm to generate sparse estimates. The final estimates of $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ may be sparse, and the estimated parameters satisfy the identifiability conditions. We remark that FIT-SSVD can be applied to the individual structure as well, if desired.

4.3. *Rank estimation.* In order to estimate (r_0, r_1, r_2), we adopt a two-step procedure. The first step is to estimate the ranks of the column centered natural parameter matrices for $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $(\boldsymbol{X}_1, \boldsymbol{X}_2)$. In order to achieve that, we devise an N-fold cross validation approach. The idea is as follows: we first randomly split the entries of a data matrix into N folds; then we withhold one fold of data and use the rest to estimate natural parameter matrices with different ranks via an alternating algorithm; finally we calculate the cross validation score corresponding to each rank by taking the average of squared Pearson residuals of the withheld data. The candidate rank with the smallest score will be selected. We remark that the

approach can flexibly accommodate a data matrix from a single non-Gaussian distribution, or a data matrix consisting of mixed variables from multiple distributions [e.g., (X_1, X_2)]. We apply the approach to X_1, X_2, and (X_1, X_2), respectively, and obtain the estimated ranks $r_1^\star$, $r_2^\star$, and $r_0^\star$.

In the second step, we solve a system of linear equations to estimate (r_0, r_1, r_2). From Model (2.1) and the identifiability conditions, we have the following relations: $r_0^\star = r_0 + r_1 + r_2$, $r_1^\star = r_0 + r_1$, and $r_2^\star = r_0 + r_2$. Therefore, the estimate of (r_0, r_1, r_2) is obtained by

$$r_0 = r_1^\star + r_2^\star - r_0^\star, \qquad r_1 = r_0^\star - r_1^\star, \qquad r_2 = r_0^\star - r_1^\star.$$

A more detailed description of the two-step rank estimation procedure and comprehensive numerical studies can be found in Section D of the Supplementary Material.

5. CAL500 music annotation. In this section, we analyze the CAL500 data. The data are publicly available at the Mulan database [Tsoumakas et al. (2011)]. The CAL500 data consist of 502 popular songs. The audio signal of each song has been analyzed via signal processing methods, and converted to 68 continuous features. The features are generally partitioned into five categories: spectral centroid, spectral flux, spectral roll-off, zero crossings, and Mel-Frequency Cepstral Coefficients (MFCC), measuring different aspects of an audio profile. In addition, each song has been manually annotated by multiple listeners. There are 174 total annotations, related to the emotion (36 variables), genre (47), usage (15), instrument (33), characteristic (27), and vocal type (16) of a song. Each song has been assigned a binary sequence of annotations based on the responses from listeners. A more detailed description can be found in Turnbull et al. (2007).

There are two data sets with matched samples but distinct data types in CAL500. The primary goal is to understand the association between the two sets of variables (i.e., acoustic features and semantic annotations), and leverage the information to achieve automatic annotation and music retrieval. The proposed GAS framework is suitable for the association analysis. In the following, we first elaborate the model fitting procedure with the CAL500 data, and then describe the annotation and retrieval performance.

5.1. *Model fitting.* Let X_1 denote the continuous acoustic features and X_2 denote the binary semantic annotations. We have $n = 502$, $p_1 = 68$, and $p_2 = 174$. Each column of X_1 has been centered and normalized to have unit standard deviation. Furthermore, we exploit SVD to estimate the standard deviation of the random noise in X_1 as σ, and scale the entire data matrix by $1/\sigma$ so that the noise has unit variance. Consequently, we model the preprocessed data X_1 by Gaussian distributions with the structured mean matrix Θ_1 in Model (2.1) and unit variance. We model the binary data matrix X_2 by Bernoulli distributions with the structured natural parameter matrix Θ_2 in Model (2.1).

We use a data-driven approach to estimate the model ranks to be $\widehat{r_0} = 3$, $\widehat{r_1} = 3$, and $\widehat{r_2} = 2$. A detailed description is provided in Section D of the Supplementary Material [Li and Gaynanova (2018)]. Subsequently, we fit Model (2.1) to the CAL500 data with the estimated ranks. We exploit the one-step approximated version of the algorithm without sparsity. The algorithm converges at high accuracy within 300 iterations, taking less than three minutes on a desktop [Intel i5 CPU (3.3 GHz) with 8 Gb RAM].

We calculate the association coefficient (3.1) based on the estimated parameters and get $\rho = 0.265$. The coefficient indicates a moderate association between the acoustic features and the semantic annotations. Furthermore, we conduct the permutation-based association test (with 1000 permutations) as described in Section 3.2. The permuted statistics roughly follow a Gaussian distribution (see Figure 1). The empirical p-value of the test is 0. Namely, the association between the acoustic features and the semantic annotations is highly statistically significant.

We further investigate the three joint loading vectors. For each loading, we sort the variables in each data source based on the loading values from large to small. In the first joint loading vector, annotations corresponding to the largest positive values include emotions such as "Soft," "Calming" and "Loving," and usage such as "Romancing." Annotations corresponding to the largest negative values include emotions such as "Aggressive" and "Angry," and genres such as "Metal Hard Rock." Namely, the first loading primarily captures the emotion of a song. The corresponding top acoustic features are the MFCCs and the zero crossings, which are known to measure the noisiness of audio signals. The second joint loading mainly characterizes the attitude of a song (e.g., "Cheerful" vs "Not Cheerful," "Danceable" vs "Not Danceable"). Music genres such as "R&B," "Soul," and "Swing"
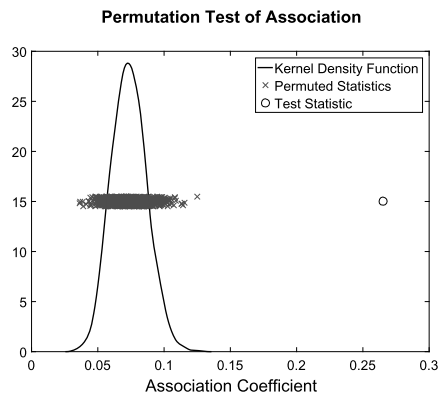

FIG. 1. *Permutation-based association test for the CAL500 data. The kernel density is estimated from 1000 permuted association coefficients. The original test statistic (black circle) and the permuted statistics (gray cross) are shown in the plot with random jitters on the y axis for the ease of visualization.*

also have large positive loading values on the cheerful side, which is quite intuitive. The corresponding top acoustic features include the MFCCs and the zero crossings, as well as the spectral centroid, which measures the "brightness" of the music texture. The third joint loading captures more subtle patterns. For annotations, genres such as "Jazz" and "Bebop" and characteristics such as "Changing Energy Level" and "Positive Feelings" have large positive values, while genres "Country," "Roots Rock," "Hip-Hop," and "Rap" have large negative values. The top acoustic features are dominated by the MFCCs.

5.2. *Automatic annotation.* Under the GAS framework, we propose the following procedure to automatically annotate a new song based on its acoustic features. Suppose we have all the model parameters, $\{\boldsymbol{\mu}_k, \boldsymbol{U}_0, \boldsymbol{V}_k, \boldsymbol{U}_k, \boldsymbol{A}_k; k = 1, 2\}$, estimated from a training data set. Given a new song with the acoustic feature vector $\boldsymbol{x}_1^\star \in \mathbb{R}^{p_1}$, we first estimate the corresponding joint and individual scores $(\boldsymbol{u}_0^{\star T}, \boldsymbol{u}_1^{\star T})^T$ by regressing $\boldsymbol{x}_1^\star - \boldsymbol{\mu}_1$ on $(\boldsymbol{V}_1, \boldsymbol{A}_1)$. Next, we extract the joint score $\boldsymbol{u}_0^\star$ and obtain an estimate of the annotation natural parameters via $\boldsymbol{\theta}_2^\star = \boldsymbol{\mu}_2 + \boldsymbol{V}_2 \boldsymbol{u}_0^\star$. Finally, we convert the natural parameters to probabilities via the entry-wise logistic transformation $\boldsymbol{\pi}^\star = \exp(\boldsymbol{\theta}_2^\star)/(1 + \exp(\boldsymbol{\theta}_2^\star))$. Consequently, each entry of $\boldsymbol{\pi}^\star$ provides the probability of the song having the corresponding annotation. In other words, $\boldsymbol{\pi}^\star$ is the induced annotation profile of the song. In practice, one could preset a threshold, and output the semantic descriptions in the vocabulary with probabilities greater than the threshold as the annotation of the song.

To compare the proposed method with existing auto-tagging approaches, we conduct a 10-fold cross validation study on the CAL500 data, similar to that in Turnbull et al. (2008). For simplicity, we select 500 out of the 502 songs in the data, and randomly partition them into 10 blocks, each having 50 songs. In each run, we use 452 songs as the training set, and test on the remaining 50 songs. To be consistent with Turnbull et al. (2008), we annotate each test song with exactly 10 annotations (the top 10 annotations with the largest probabilities in $\boldsymbol{\pi}^\star$ according to our method).

The annotation performance is assessed by the mean per-word precision and recall. More specifically, for each annotation, let t_{GT} be the number of songs in the test set that have the annotation in the human-generated "ground truth"; let t_{A} be the number of songs that are annotated with the tag by a method; let t_{TP} be the number of "true positives" that have the tag both in the ground truth and in the automatic annotation prediction. The per-word precision is defined as $t_{\text{TP}}/t_{\text{A}}$, and the per-word recall is $t_{\text{TP}}/t_{\text{GT}}$. The mean per-word precision and recall are calculated by averaging the values across different tags in each cross validation run. Annotations with undefined precision or recall are omitted when calculating the mean.

TABLE 2
*The CAL*500 *automatic annotation results. The mean and standard deviation* (*in parenthesis*) *for mean per-word precision* (*"Precision"*) *and mean per-word recall* (*"Recall"*) *across* 10 *cross validation runs are presented. The best results are bold-faced*

Method	Precision	Recall
Random	0.144 (0.004)	0.064 (0.002)
UpperBnd	0.712 (0.007)	0.375 (0.006)
MixHier	0.265 (0.007)	**0.158** (0.006)
Autotagger	0.312 (0.060)	0.153 (0.015)
Proposed	**0.438** (0.051)	0.078 (0.007)

We compare the proposed method with the MixHier method [Turnbull et al. (2008)] and the Autotagger method [Bertin-Mahieux et al. (2008)]. We also consider two baseline methods, a "Random" lower bound and an empirical upper bound (denoted by "UpperBnd"), for precision and recall, as discussed in Turnbull et al. (2008). Loosely speaking, the Random approach randomly selects 10 annotations for each test song based on the observed tag frequencies, and mimics a random guessing procedure. The UpperBnd approach serves as the best-case scenario. It uses the ground truth to annotate test songs, and randomly adds or removes tags to meet the ten-annotation requirement. The mean and standard deviation of the mean per-word precision and recall for different methods from the 10-fold cross validation are presented in Table 2.

Overall, all three methods are significantly better than random guessing, but considerably worse than the empirical upper bounds. The suboptimal results may be justified by the moderate association between the acoustic features and the semantic annotations (see Section 5.1). Namely, only a moderate amount of information in the annotations can be explained by the existing acoustic features. Thus, to further improve the automatic annotation performance, more comprehensive characterization of the audio profile may be needed.

Although a good balance of precision and recall is desired, it has been argued that precision is more relevant for recommender systems [Herlocker, Konstan and Riedl (2000)]. The proposed method has the best precision among all three methods. Thus, it may provide an effective approach for auto-tagging. The relatively low recall may be due to the small number of predicted annotations (i.e., 10) per song. We further increase the number of words used to characterize a song to 20, and redo the analysis. As a result, we get a recall rate of 0.154 with standard deviation 0.015, which is comparable to the competing methods, and a precision rate of 0.330 with standard deviation 0.036, which is still superior to the competing methods. We further investigate the complete annotation profile of each song using the proposed method. Figure 2 shows four randomly selected examples. The
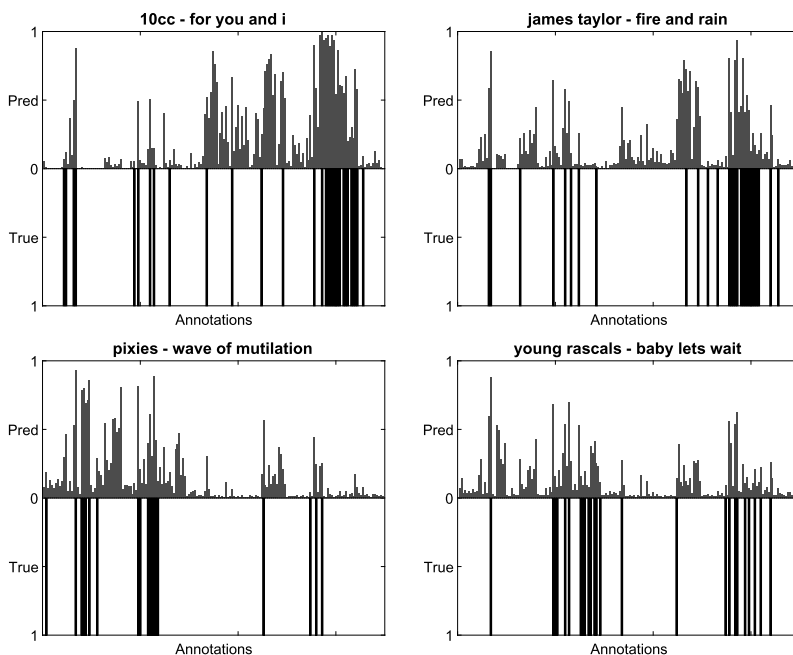
FIG. 2. *The CAL500 automatic annotation results. Each plot corresponds to a song. In each plot, the top gray bars provide the predicted annotation profile; the lower black bars correspond to the true annotations. The annotations are ordered for visualization convenience.*

top and bottom bars in each plot correspond to the estimated and true annotation profiles. We particularly order the annotations for visualization convenience. The proposed method produces sensible results. It captures the majority of the true annotations with large probabilities, and has much richer patterns. Whether the additional annotations with high probabilities are false positives or missing tags due to the well-known "human bias" issue in music tagging [Ellis et al. (2002)] remains an open question.

5.3. *Music retrieval.*   We also investigate music retrieval using the proposed framework. We remark that finding songs based on a small set of annotations is relatively easy. One could simply filter the songs in the database by the given tags, and output those satisfying all the requirements. Thus it is not our primary interest here. Instead, we focus on retrieving songs according to a more complicated query consisting of multiple tags.

Similar to automatic annotation, we propose the following procedure for music retrieval based on a given annotation list. Suppose the model parameters in (2.1) have been estimated. For any given query, we first convert it to a binary vector $x_2^\star$ using the semantic annotation library. Then, we regress $x_2^\star$ on (V_2, A_2) using a logistic regression with offset μ_2, and obtain the estimate of the joint and indi-
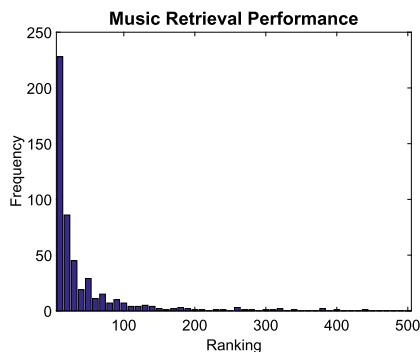
FIG. 3. *The CAL500 music retrieval result. The histogram of the reference song rankings across different music retrieval requests.*

vidual scores $u_0^\star$ and $u_2^\star$. Next, we calculate the Mahalanobis distances between the estimated score vector $(u_0^{\star T}, u_2^{\star T})^T$ and the score vectors corresponding to the songs in the database. The covariance matrix used in the Mahalanobis distance is estimated from the model parameter (U_0, U_2). Finally, we sort the distances in an ascending order. As a result, we obtain an ordered list with highest recommendation on the top.

To validate the procedure, we apply it to the CAL500 data. We use the annotation profile of each song as a query. For each query, we record the ranking of the reference song (also contained in the database) in the output recommendation list. Figure 3 shows the histogram of the rankings across the 502 requests. As desired, most of the time, the reference song is among the top of the recommendation list. Perhaps what's more interesting are the top choices other than the reference song in each request. They are the most similar songs to the reference song in the database according to the annotation query. For instance, for the song "For You and I" by 10cc, the top recommendations include "God Bless the Child" by Billie Holiday, "Suzanne" by Leonard Cohen, and "Postcard Blues" by Cowboy Junkies. Without "ground truth" of the true rankings, however, further validation of the music retrieval performance remains an open question [Ellis et al. (2002)].

6. Simulation study. In this section, we conduct comprehensive simulation studies to compare the proposed method with existing ones. We consider several versions of the method: the double-iterative version (denoted by "iter-GAS") as described in Algorithm 1, the one-step version ("GAS") as described in Section C of the Supplementary Material [Li and Gaynanova (2018)], and the one-step with sparsity version ("sGAS") as described in Section 4.2. In addition, we also consider an ad hoc competing method derived from EPCA [Collins, Dasgupta and Schapire (2001)] and JIVE [Lock et al. (2013)], where we first estimate a low-rank individual natural parameter matrix for each data set via EPCA, and then apply

JIVE to the two estimated matrices. We denote the ad hoc approach by EPCA-JIVE.

We generate data from Model (2.1), and apply different methods to estimate model parameters. To avoid complication, we set the joint and individual ranks for the GAS methods to be the true ranks. In Section G of the Supplementary Material [Li and Gaynanova (2018)], we further investigate the effect of rank misspecification on the performance. For the EPCA-JIVE method, in the EPCA step, we set the rank of each individual natural parameter matrix to be a large number (much larger than the true rank) in order to avoid information loss. In particular, for Gaussian data, we use the full rank, or equivalently, the original data. In the JIVE step, we use the true joint and individual ranks. The assessment of the rank estimation procedure is conducted separately in Section D.3 of the Supplementary Material [Li and Gaynanova (2018)].

6.1. *Setting.* We set the sample size to be $n = 200$, and the dimensions of both data sets to be $p_1 = p_2 = 120$. The joint and individual ranks of the column-centered natural parameter matrices are $r_0 = r_1 = r_2 = 2$. The scores in (U_0, U_1, U_2) are filled with random numbers generated from a uniform distribution between -0.5 to 0.5 [i.e., Unif$(-0.5, 0.5)$], and normalized via the Gram–Schmidt process to have orthonormal columns. We particularly consider four settings of the natural parameters, and perform 100 simulation runs for each with the same underlying parameters.

- **Setting 1 (Gaussian–Gaussian)**: The joint loadings $(V_1^T, V_2^T)^T$ are generated in a similar way to the scores: filled with uniform random numbers and normalized to have orthonormal columns. The respective individual loadings A_1 and A_2 are similarly generated to satisfy the identifiability conditions. We set the singular values of the joint structure to be $(180, 140)$, and of the individual structure to be $(120, 100)$ and $(100, 80)$. All singular values are absorbed into the scores. The intercepts μ_1 and μ_2 are filled with Unif$(-0.5, 0.5)$.
- **Setting 2 (Gaussian–Bernoulli)**: The loadings are generated similarly to Setting 1, except that V_1 (Gaussian) and V_2 (Bernoulli) are initially filled with Unif$(-0.5, 0.5)$ and Unif$(-1, 1)$ before the normalization. The singular values of the joint structure are $(240, 220)$ and those for the individual structure are $(90, 80)$ and $(200, 180)$. The intercept is filled with Unif$(-0.5, 0.5)$.
- **Setting 3 (Gaussian–Poisson)**: The loadings are generated similarly to Setting 1, except that V_1 (Gaussian) and V_2 (Poisson) are initially filled with Unif$(-0.5, 0.5)$ and Unif$(-0.25, 0.25)$. The singular values are $(80, 40)$ (joint), $(60, 40)$ (Gaussian individual), and $(20, 16)$ (Poisson individual). The intercept terms μ_1 and μ_2 are filled with Unif$(-0.5, 0.5)$ and Unif$(2, 3)$ respectively.
- **Setting 4 (Bernoulli–Poisson)**: The loadings are generated similarly to Setting 1, except that V_1 (Bernoulli) and V_2 (Poisson) are initially filled with Unif$(-5, 5)$ and Unif$(-0.5, 0.5)$ respectively. The singular values are $(180,140)$

(joint), (200, 160) (Bernoulli individual), and (12, 10) (Poisson individual). The intercept terms $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are filled with Unif$(-0.5, 0.5)$ and Unif$(2, 3)$ respectively.

Once the natural parameters are fixed, the observed data are generated independently from corresponding distributions. In particular, for Gaussian random numbers, we set the variance to be one.

We remark that for Bernoulli distribution, the scale of the natural parameters needs to be relatively large in order to have a detectable signal. Hence we purposely increase the corresponding singular values and the relative loading scales for the Bernoulli distribution in **Setting 2 and 4**. For Poisson distribution, due to the asymmetry of the canonical link function, the natural parameters are typically skewed towards positive values. To mimic reality, we set the intercept term for the Poisson distribution to be positive in **Setting 3 and 4**.

We also consider the settings where the joint loadings are sparse. As the results for sparse settings are qualitatively similar to the results in dense settings, we refer the reader to Section F of the Supplementary Material [Li and Gaynanova (2018)].

6.2. *Result.*   We compare GAS, iter-GAS, and EPCA-JIVE on the nonsparse simulation settings. Each method is applied to the simulated data to estimate the model parameters. We evaluate the loading estimation accuracy by the maximum principal angle [Björck and Golub (1973)] between the subspaces spanned by the estimated and the true loading vectors. We consider the angles for the joint loadings $\angle(\boldsymbol{V}_0, \widehat{\boldsymbol{V}}_0)$ [where $\boldsymbol{V}_0 = (\boldsymbol{V}_1^T, \boldsymbol{V}_2^T)^T$] and for separate individual loadings $\angle(\boldsymbol{A}_k, \widehat{\boldsymbol{A}}_k)$ $(k = 1, 2)$, respectively. We assess the estimation accuracy of different model parameters (i.e., the intercept, the joint, and the individual structure) by the Frobenius norm of the difference between the true and the estimated values. In particular, we calculate the following quantities $(k = 1, 2)$:

$$\text{Norm}_{\text{avg}} = \|\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}_k}\|_{\mathbb{F}},$$

$$\text{Norm}_{\text{jnt}} = \|\boldsymbol{U}_0 \boldsymbol{V}_k^T - \widehat{\boldsymbol{U}_0} \widehat{\boldsymbol{V}}_k^T\|_{\mathbb{F}},$$

$$\text{Norm}_{\text{ind}} = \|\boldsymbol{U}_k \boldsymbol{A}_k^T - \widehat{\boldsymbol{U}_k} \widehat{\boldsymbol{A}}_k^T\|_{\mathbb{F}},$$

where $\|\cdot\|_{\mathbb{F}}$ represents the Frobenius norm. Moreover, we also calculate the Frobenius loss of the overall natural parameter estimates $\text{Norm}_{\boldsymbol{\Theta}} = \|\boldsymbol{\Theta}_k - \widehat{\boldsymbol{\Theta}_k}\|_{\mathbb{F}}$. In addition, we compare the model fitting times for different methods. The results are summarized in Table 3.

We observe that under **Setting 1** where the two data sets are both Gaussian, all three methods have very similar performances. In particular, GAS and iter-GAS are identical because the IRLS algorithm degenerates to the ordinary least squares under the Gaussian assumption. Model (2.1) coincides with the JIVE model in this setting, and thus GAS provides an alternative way of fitting the JIVE model. In **Setting 2** where the distributions are Gaussian and Bernoulli, the GAS method

TABLE 3

Simulation results based on 100 simulation runs in each setting. The median and median absolute deviation (in parenthesis) of each criterion for different methods across different settings are presented. For each method, Norm$_{avg}$, Norm$_{jnt}$, Norm$_{ind}$, Norm$_\Theta$ and $\angle(A_k, \widehat{A}_k)$ are evaluated and compared per data set; $\angle(V_0, \widehat{V}_0)$ is evaluated across two data sets. The best results are highlighted in bold

		GAS		iter-GAS		EPCA-JIVE	
		Data 1	**Data 2**	**Data 1**	**Data 2**	**Data 1**	**Data 2**
Setting 1	$\|\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}_k}\|_{\mathbb{F}}$	**0.78** (0.03)	**0.77** (0.04)	**0.78** (0.03)	**0.77** (0.04)	**0.78** (0.03)	**0.77** (0.04)
	$\|\boldsymbol{U}_0\boldsymbol{V}_k^T - \widehat{\boldsymbol{U}_0}\widehat{\boldsymbol{V}_k}^T\|_{\mathbb{F}}$	**21.32** (0.43)	**21.15** (0.41)	**21.32** (0.43)	**21.15** (0.41)	21.33 (0.42)	**21.15** (0.41)
	$\|\boldsymbol{U}_k\boldsymbol{A}_k^T - \widehat{\boldsymbol{U}_k}\widehat{\boldsymbol{A}_k}^T\|_{\mathbb{F}}$	**25.39** (0.51)	**25.65** (0.53)	**25.39** (0.51)	**25.65** (0.53)	**25.39** (0.51)	**25.65** (0.53)
	$\|\boldsymbol{\Theta}_k - \widehat{\boldsymbol{\Theta}_k}\|_{\mathbb{F}}$	**34.61** (0.39)	**34.58** (0.49)	**34.61** (0.39)	**34.58** (0.49)	**34.61** (0.40)	**34.58** (0.49)
	$\angle(\boldsymbol{A}_k, \widehat{\boldsymbol{A}_k})$	**6.27** (0.27)	**7.96** (0.30)	**6.27** (0.27)	**7.96** (0.30)	**6.27** (0.26)	**7.96** (0.30)
	$\angle(\boldsymbol{V}_0, \widehat{\boldsymbol{V}_0})$	**6.36** (0.20)		**6.36** (0.20)		**6.36** (0.20)	
	Time (sec)	10.04 (0.82)		44.78 (3.27)		**0.51** (0.01)	
Setting 2	$\|\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}_k}\|_{\mathbb{F}}$	**0.78** (0.04)	2.54 (0.10)	**0.78** (0.03)	**1.96** (0.10)	**0.78** (0.04)	2.59 (0.10)
	$\|\boldsymbol{U}_0\boldsymbol{V}_k^T - \widehat{\boldsymbol{U}_0}\widehat{\boldsymbol{V}_k}^T\|_{\mathbb{F}}$	**23.69** (0.45)	**89.36** (5.63)	42.79 (0.56)	128.98 (1.00)	25.15 (0.48)	185.51 (1.07)
	$\|\boldsymbol{U}_k\boldsymbol{A}_k^T - \widehat{\boldsymbol{U}_k}\widehat{\boldsymbol{A}_k}^T\|_{\mathbb{F}}$	**26.00** (0.40)	**110.89** (5.30)	26.01 (0.45)	133.88 (1.04)	26.11 (0.44)	174.32 (1.04)
	$\|\boldsymbol{\Theta}_k - \widehat{\boldsymbol{\Theta}_k}\|_{\mathbb{F}}$	**36.08** (0.45)	**146.86** (7.47)	50.80 (0.45)	187.77 (0.96)	37.09 (0.48)	257.07 (1.14)
	$\angle(\boldsymbol{A}_k, \widehat{\boldsymbol{A}_k})$	**8.18** (0.40)	14.47 (0.69)	8.20 (0.38)	**13.95** (0.60)	8.24 (0.38)	22.03 (0.99)
	$\angle(\boldsymbol{V}_0, \widehat{\boldsymbol{V}_0})$	12.96 (0.79)		**12.70** (0.40)		29.46 (0.43)	
	Time (sec)	**10.94** (1.36)		55.13 (6.39)		43.21 (3.71)	
Setting 3	$\|\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}_k}\|_{\mathbb{F}}$	**0.77** (0.03)	**0.23** (0.01)	**0.77** (0.03)	**0.23** (0.01)	**0.77** (0.03)	0.25 (0.01)
	$\|\boldsymbol{U}_0\boldsymbol{V}_k^T - \widehat{\boldsymbol{U}_0}\widehat{\boldsymbol{V}_k}^T\|_{\mathbb{F}}$	**18.65** (0.49)	**6.68** (0.14)	**18.65** (0.49)	6.69 (0.14)	76.32 (4.29)	22.16 (3.58)
	$\|\boldsymbol{U}_k\boldsymbol{A}_k^T - \widehat{\boldsymbol{U}_k}\widehat{\boldsymbol{A}_k}^T\|_{\mathbb{F}}$	**26.31** (0.53)	**7.16** (0.16)	**26.31** (0.53)	**7.16** (0.16)	76.63 (4.00)	28.22 (3.04)
	$\|\boldsymbol{\Theta}_k - \widehat{\boldsymbol{\Theta}_k}\|_{\mathbb{F}}$	33.98 (0.45)	**10.15** (0.13)	**33.97** (0.45)	**10.15** (0.13)	37.86 (0.46)	18.93 (0.13)
	$\angle(\boldsymbol{A}_k, \widehat{\boldsymbol{A}_k})$	**15.96** (0.77)	**11.49** (0.55)	**15.96** (0.77)	**11.49** (0.55)	84.31 (4.17)	88.51 (1.00)
	$\angle(\boldsymbol{V}_0, \widehat{\boldsymbol{V}_0})$	**16.28** (0.60)		**16.28** (0.60)		85.68 (3.21)	
	Time (sec)	**23.10** (1.28)		111.32 (6.58)		54.15 (6.59)	

TABLE 3
(*Continued*)

		GAS		iter-GAS		EPCA-JIVE	
		Data 1	**Data 2**	**Data 1**	**Data 2**	**Data 1**	**Data 2**
	$\|\boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}_k}\|_{\mathbb{F}}$	2.36 (0.12)	**0.23** (0.01)	**1.87** (0.08)	**0.23** (0.01)	2.48 (0.07)	0.24 (0.01)
	$\|\boldsymbol{U}_0\boldsymbol{V}_k^T - \widehat{\boldsymbol{U}_0}\widehat{\boldsymbol{V}_k}^T\|_{\mathbb{F}}$	**82.99** (4.23)	**6.17** (0.11)	101.71 (1.16)	7.81 (0.17)	203.54 (3.13)	16.59 (0.89)
Setting 4	$\|\boldsymbol{U}_k\boldsymbol{A}_k^T - \widehat{\boldsymbol{U}_k}\widehat{\boldsymbol{A}_k}^T\|_{\mathbb{F}}$	**106.96** (5.51)	**7.50** (0.15)	119.11 (1.09)	7.54 (0.15)	233.41 (0.77)	20.11 (0.88)
	$\|\boldsymbol{\Theta}_k - \widehat{\boldsymbol{\Theta}_k}\|_{\mathbb{F}}$	**138.99** (5.22)	**10.17** (0.14)	157.89 (1.22)	11.27 (0.15)	218.95 (1.21)	13.96 (0.14)
	$\angle(\boldsymbol{A}_k, \widehat{\boldsymbol{A}_k})$	14.37 (0.84)	**18.88** (0.94)	**13.29** (0.74)	18.97 (0.92)	86.86 (1.96)	88.57 (0.90)
	$\angle(\boldsymbol{V}_0, \widehat{\boldsymbol{V}_0})$	15.39 (1.02)		**14.98** (0.78)		87.59 (1.64)	
	Time (sec)	**7.42** (0.63)		35.53 (3.18)		81.13 (5.01)	

is generally the best (except for the mean structure and loading estimation in the second data set). For Bernoulli distributions, sometimes the maximum likelihood of EPCA and iter-GAS is reached at infinity, posing a convergence issue to both methods. The same issue has been pointed out in Collins, Dasgupta and Schapire (2001). As a remedy, we introduce a small ridge penalty to the GLM likelihood functions. This allows the algorithm to converge to a finite value. However, the resulting estimates are biased and shrunk towards zero. See Section E of the Supplementary Material [Li and Gaynanova (2018)] for more details. We remark that the one-step approximation algorithm is more robust against the convergence issue, and typically does not require such a penalty. Consequently, the estimates are more accurate. In **Setting 3** where the distributions are Gaussian and Poisson, GAS and iter-GAS have similar results, both outperforming the EPCA-JIVE method. In **Setting 4** where the distributions are Bernoulli and Poisson, again, GAS is generally among the best in almost all aspects, followed by iter-GAS. Both provide more accurate estimates than EPCA-JIVE. In terms of the computational cost, the one-step GAS method is always more efficient than the iterative GAS method. Both outperform the ad hoc approach except for the Gaussian case.

As suggested by a referee, we also investigate the performance of the GAS method in high-dimensional settings. We focus on **Setting 3** and consider two variants with dimensions $p_1 = p_2 = 200$ and $p_1 = p_2 = 300$, respectively. We keep the signal-to-noise ratio constant as the dimensions increase. Analysis results show that the estimation accuracy further improves with increasing dimensions due to the "blessing of dimensionality" [Li et al. (2018)], demonstrating the efficacy of the GAS method in high-dimensional settings. More details can be found in Section G of the Supplementary Material [Li and Gaynanova (2018)].

In addition, we also study the proposed method in the situation where ranks are misspecified. Results show that the estimation of underlying natural parameter matrices, loading subspaces, and association coefficients is very robust against rank misspecification. More details can be found in Section H of the Supplementary Material [Li and Gaynanova (2018)].

7. Discussion. In this paper, we develop a generalized association study framework for estimating the dependency structure and testing the significance of association between two heterogeneous data sets. We analyze the CAL500 music annotation data with the proposed method, and identify a statistically significant but moderate association between the acoustic features and the semantic annotations. By leveraging the information in both data sets, we develop new auto-tagging and music retrieval methods that show superior precision performance over existing approaches. As such, they may serve as useful tools for recommender systems.

There are a few interesting directions for future research. First, for the music annotation study, it is compelling to investigate what additional audio features may significantly enhance the association with the semantic annotations and improve the auto-tagging performance. Second, from a methodological point of view, the

proposed framework may be extended to over-dispersed distributions and/or to more than two data sets. How to simultaneously estimate dispersion parameters is an open question. Third, the application of the proposed methods to other areas such as multi-omics studies is open and promising.

SUPPLEMENTARY MATERIAL

Supplementary Material for A General Framework for Association Analysis of Heterogeneous Data (DOI: 10.1214/17-AOAS1127SUPP; .pdf). We provide proofs, technical details of the algorithm, a detailed description of the rank estimation procedure, and additional simulation results in the supplementary material.

REFERENCES

BACH, F. R. and JORDAN, M. I. (2005). A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Dept. Statistics, Univ. California, Berkeley, Berkeley, CA.

BARRINGTON, L., CHAN, A., TURNBULL, D. and LANCKRIET, G. (2007). Audio information retrieval using semantic similarity. In *International Conference on Acoustics*, *Speech and Signal Processing* **2** 725–728. IEEE, New York.

BERTIN-MAHIEUX, T., ECK, D., MAILLET, F. and LAMERE, P. (2008). Autotagger: A model for predicting social tags from acoustic features on large music databases. *J. New Music Res*. **37** 115–135.

BJÖRCK, K. and GOLUB, G. H. (1973). Numerical methods for computing angles between linear subspaces. *Math. Comp*. **27** 579–594.

BROWNE, M. W. (1979). The maximum-likelihood solution in inter-battery factor analysis. *Br. J. Math. Stat. Psychol*. **32** 75–86. MR0553146

CHAUDHURI, K., KAKADE, S. M., LIVESCU, K. and SRIDHARAN, K. (2009). Multi-view clustering via canonical correlation analysis. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* 129–136. ACM, New York.

CHEN, X. and LIU, H. (2012). An efficient optimization algorithm for structured sparse cca, with applications to eqtl mapping. *Stat. Biosci*. **4** 3–26.

CHEN, M., GAO, C., REN, Z. and ZHOU, H. H. (2013). Sparse cca via precision adjusted iterative thresholding. ArXiv preprint. Available at arXiv:1311.6186.

CHENG, J., LI, T., LEVINA, E. and ZHU, J. (2017). High-dimensional mixed graphical models. *J. Comput. Graph. Statist*. **26** 367–378. MR3640193

COLLINS, M., DASGUPTA, S. and SCHAPIRE, R. E. (2001). A generalization of principal components analysis to the exponential family. In *NIPS'*01: *Proceedings of the* 14*th International Conference on Neural Information Processing Systems*: *Natural and Synthetic* 617–624. MIT Press, Cambridge, MA.

ELLIS, D. P., WHITMAN, B., BERENZWEIG, A. and LAWRENCE, S. (2002). The quest for ground truth in musical artist similarity. In *ISMIR* 2002 *Conference Proceedings*: *Third International Conference on Music Information Retrieval*: *October* 13–17, 2002, *IRCAM-Centre Pompidou*, *Paris*, *France*.

GOLDSMITH, J., ZIPUNNIKOV, V. and SCHRACK, J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics* **71** 344–353.

GOLUB, G. H. and VAN LOAN, C. F. (2013). *Matrix Computations*, 4th ed. *Johns Hopkins Studies in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR3024913

GOTO, M. and HIRATA, K. (2004). Recent studies on music information processing. *Acoust. Sci. Technol.* **25** 419–425.

HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity*: The *Lasso and Generalizations*. CRC Press, Boca Raton, FL.

HERLOCKER, J. L., KONSTAN, J. A. and RIEDL, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the* 2000 *ACM Conference on Computer Supported Cooperative Work* 241–250. ACM, New York.

HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika* **28** 321–377.

JIA, Y., SALZMANN, M. and DARRELL, T. (2010). Factorized latent spaces with structured sparsity. *Adv. Neural Inf. Process. Syst.* 982–990.

JOHNSON, N. L., KOTZ, S. and BALAKRISHNAN, N. (1997). *Discrete Multivariate Distributions* **165**. Wiley, New York.

KLAMI, A., VIRTANEN, S. and KASKI, S. (2010). Bayesian exponential family projections for coupled data sources. In *The Twenty-Sixth Conference on Uncertainty in Artificial Intelligence* 286–293. AUAI Press.

KLAMI, A., VIRTANEN, S. and KASKI, S. (2013). Bayesian canonical correlation analysis. *J. Mach. Learn. Res.* **14** 965–1003.

LANDGRAF, A. J. and LEE, Y. (2015). Generalized principal component analysis: Projection of saturated model parameters. Technical Report 892, Department of Statistics, Ohio State Univ.

LI, G. and GAYNANOVA, I. (2018). Supplement to "A general framework for association analysis of heterogeneous data." DOI:10.1214/17-AOAS1127SUPP.

LI, Q., CHENG, G., FAN, J. and WANG, Y. (2018). Embracing the blessing of dimensionality in factor models. *J. Amer. Statist. Assoc.* **113** 380–389. MR3803472

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. *Ann. Appl. Stat.* **7** 523–542.

LOGAN, B. (2000). Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval* (*ISMIR*).

LUO, C., LIU, J., DEY, D. K. and CHEN, K. (2016). Canonical variate regression. *Biostatistics* **17** 468–483.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, London. [Second edition of MR0727836.] MR3223057

SHE, Y. (2013). Reduced rank vector generalized linear models for feature extraction. *Stat. Interface* **6** 197–209.

TRYGG, J. and WOLD, S. (2003). O2–PLS, a two-block (X–Y) latent variable regression (LVR) method with an integral OSC filter. *J. Chemom.* **17** 53–64.

TSOUMAKAS, G., SPYROMITROS-XIOUFIS, E., VILCEK, J. and VLAHAVAS, I. (2011). Mulan: A Java library for multi-label learning. *J. Mach. Learn. Res.* **12** 2411–2414.

TUCKER, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika* **23** 111–136. MR0099737

TURNBULL, D., BARRINGTON, L., TORRES, D. and LANCKRIET, G. (2007). Towards musical query-by-semantic-description using the cal500 data set. In *Proceedings of the* 30*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 439–446. ACM, New York.

TURNBULL, D., BARRINGTON, L., TORRES, D. and LANCKRIET, G. (2008). Semantic annotation and retrieval of music and sound effects. *IEEE/ACM Trans. Audio Speech Lang. Process.* **16** 467–476.

VIRTANEN, S., KLAMI, A. and KASKI, S. (2011). Bayesian cca via group sparsity. In *Proceedings of the* 28*th International Conference on Machine Learning* (*ICML* 2011) 457–464. ACM, New York.

WESTERHUIS, J. A., KOURTI, T. and MACGREGOR, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemom.* **12** 301–321.

WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 513–534.

YANG, D., MA, Z. and BUJA, A. (2014). A sparse singular value decomposition method for high-dimensional data. *J. Comput. Graph. Statist.* **23** 923–942. MR3270704

YANG, Z., NING, Y. and LIU, H. (2014). On semiparametric exponential family graphical models. ArXiv preprint. Available at arXiv:1412.8697.

ZHOU, G., CICHOCKI, A., ZHANG, Y. and MANDIC, D. P. (2016a). Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Trans. Neural Netw. Learn. Syst.* **27** 2426–2439.

ZHOU, G., ZHAO, Q., ZHANG, Y., ADALI, T., XIE, S. and CICHOCKI, A. (2016b). Linked component analysis from matrices to high-order tensors: Applications to biomedical data. *Proc. IEEE* **104** 310–331.

ZOH, R. S., MALLICK, B., IVANOV, I., BALADANDAYUTHAPANI, V., MANYAM, G., CHAPKIN, R. S., LAMPE, J. W. and CARROLL, R. J. (2016). PCAN: Probabilistic correlation analysis of two non-normal data sets. *Biometrics* **72** 1358–1368. MR3591620

DEPARTMENT OF BIOSTATISTICS
MAILMAN SCHOOL OF PUBLIC HEALTH
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK 10032
USA
E-MAIL: gl2521@cumc.columbia.edu

DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843
USA
E-MAIL: irinag@stat.tamu.edu