# Bayesian Multiview Factor Modeling for Integrating and Analyzing Heterogeneous Clinical Data

**Wenzhao Lian, Piyush Rai, Esther Salazar, Lawrence Carin**
ECE Department, Duke University
Durham, NC 27708
{wenzhao.lian,piyush.rai,esther.salazar,lcarin}@duke.edu

## 1 Introduction

Modern-day clinical data analysis problems are typically characterized by heterogeneous data having multiple representations or *views*. Consider a problem from cognitive neuroscience [16], where data collected from a set of individuals may include their *ordinal-valued* responses on multiple questionnaires, their *real-valued* measurements from fMRI or EEG, and one or more sets of their *pairwise similarities* (e.g., computed using SNP measurements or other sources). Each representation is essentially a "view" of the data and the goal here is to integrate these diverse views to uncover the *latent* traits (factors) of these individuals, or learn a classifier for predicting certain psychopathological conditions in these individuals, or predict the missing data in one or more views (e.g., predicting missing fMRI data, leveraging information from the other views).

In this paper, we present framework to integrate such heterogeneous data (real-,binary-,ordinal-valued *feature matrices* and/or *similarity matrices*), with potentially missing data in each data source, and learn latent factors describing each individual. In addition, the proposed framework is also able to identify global as well as view-specific latent factors, and the correlations among the data sources. Our framework also admits a seamless adaptation for classification, e.g., when the eventual goal is to predict certain psychopathological conditions for the individuals under a clinical study, based on the data from all the views. Our framework also consists of an efficient, variational inference algorithm which, in addition to being scalable for large data sets, is appealing in its own right by providing a principled way to learn the *cutpoints* for data in the ordinal-valued views, which can be useful for the general problem of modeling ordinal-valued data (e.g., questionnaire responses).

## 2 A Generative Framework for Heterogeneous Multiview Data

We now describe our basic framework for **M**ultiview **L**earning with **F**eatures and **S**imilarities (abbreviated henceforth as MLFS), for modeling heterogeneous multiview data. We assume the data consist of $N$ objects (e.g., individuals in a clinical study) having a total of $M$ feature-based and/or similarity-based views. Of the $M = M_1 + M_2 + M_3$ views, the first $M_1$ are assumed to be ordinal feature matrices $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(M_1)}$ (binary feature matrix is a special case), the next $M_2$ views are assumed to be real-valued feature matrices $\boldsymbol{X}^{(M_1+1)}, \ldots, \boldsymbol{X}^{(M_1+M_2)}$, and the remaining $M_3$ views are assumed to be real-valued similarity matrices $\boldsymbol{X}^{(M_1+M_2+1)}, \ldots, \boldsymbol{X}^{(M_1+M_2+M_3)}$. One or more of these matrices may have missing data (randomly missing entries or randomly missing entire rows and/or columns). For a feature-based view, $\boldsymbol{X}^{(m)}$ denotes a feature matrix of size $N \times D_m$; for a similarity-based view, $\boldsymbol{X}^{(m)}$ denotes a similarity matrix of size $N \times N$. We assume the data $\boldsymbol{X}^{(m)}$ in each feature/similarity-based view are generated from a latent real-valued matrix $\boldsymbol{U}^{(m)} = [\boldsymbol{U}_1^{(m)}; \ldots; \boldsymbol{U}_N^{(m)}] \in \mathbb{R}^{N \times K_m}$, where $\boldsymbol{U}_i^{(m)}, i = 1, \ldots, N$ are assumed to be row vectors.

**Feature-based Views:** The $N \times D_m$ feature matrix $\boldsymbol{X}^{(m)}$ for view $m$ is generated, via a link-function $f_m$, from a real-valued matrix $\boldsymbol{U}^{(m)}$ of the same size (thus $K_m = D_m$). Therefore, $\boldsymbol{X}_{id}^{(m)} = f_m(\boldsymbol{U}_{id}^{(m)})$ where $i$ indexes the $i$-th object and $d$ indexes the $d$-th feature. For real-valued data, the link-function is identity, so $\boldsymbol{X}_{id}^{(m)} = \boldsymbol{U}_{id}^{(m)}$. For ordinal data in view $m$ having $L_m$ levels $(1, \cdots, L_m)$, $\boldsymbol{X}_{id}^{(m)} = l$ if $g_{l-1}^m < \boldsymbol{U}_{id}^{(m)} < g_l^m$, with *cutpoints* $-G = g_0^m < g_1^m < g_2^m < \ldots < g_{L_m-1}^m < g_{L_m}^m = +G$. Because the cutpoints contain information indicating relative frequencies of ordinal outcomes in each view, we will learn them as part of our variational inference procedure.

**Similarity-based Views:** The $N \times N$ similarity matrix $\boldsymbol{X}^{(m)}$ of view $m$ is generated as $\boldsymbol{X}_{ij}^{(m)} \sim \mathcal{N}or(\boldsymbol{U}_i^{(m)}{\boldsymbol{U}_j^{(m)}}^\top, \tau_m^{-1})$ where $\boldsymbol{X}_{ij}^{(m)}$ denotes the pairwise similarity between objects $i$ and $j$ in view $m$. In this work, we consider symmetric similarity matrices and thus only model $\boldsymbol{X}_{ij}^{(m)}, i < j$, but the model can be naturally extended to asymmetric cases. In this case, $\boldsymbol{U}^{(m)} \in \mathbb{R}^{N \times K_m}$ is akin to a low-rank approximation of the similarity matrix $\boldsymbol{X}^{(m)}$ ($K_m < N$).

Although the view-specific latent matrices $\{\boldsymbol{U}^{(m)}\}_{m=1}^M$ have different meanings (and play different roles) in feature-based and similarity-based views, in both cases there exists a mapping from $\boldsymbol{U}^{(m)}$ to the observed data $\boldsymbol{X}^{(m)}$. We wish to extract and summarize the information from all these view-specific latent matrices $\{\boldsymbol{U}^{(m)}\}_{m=1}^M$ to obtain a *global* latent representation of the data, and use it for tasks such as classification or clustering. To do so, we assume the view-specific latent matrices $\{\boldsymbol{U}^{(m)}\}_{m=1}^M$ as being generated from a *shared* real-valued latent factor matrix $\boldsymbol{V} = [\boldsymbol{V}_1; \ldots; \boldsymbol{V}_N]$ of size $N \times R$ (where $R$ denotes the number of latent factors) with view-specific *sparse* factor loading matrices $\boldsymbol{W} = \{\boldsymbol{W}^{(m)}\}_{m=1}^M$: $\boldsymbol{U}_i^{(m)} \sim \mathcal{N}or(\boldsymbol{V}_i\boldsymbol{W}^{(m)}, \gamma_m^{-1}\boldsymbol{I})$, where $\boldsymbol{W}^{(m)} \in \mathbb{R}^{R \times K_m}$.

Since different views may capture different aspects, we impose a structured-sparsity prior in the factor loading matrices $\boldsymbol{W} = \{\boldsymbol{W}^{(m)}\}_{m=1}^M$ of all the views, such that some of the rows in these matrices share the same support for non-zero entries whereas some rows are non-zero only for a subset of these matrices. Figure 1 summarizes our basic framework.

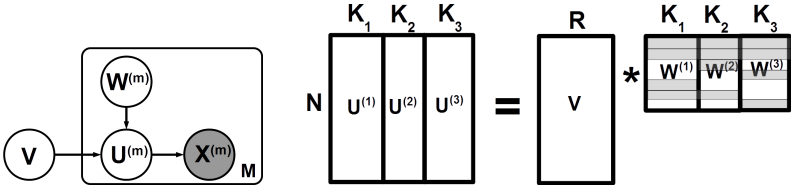


Figure 1: Left: plate notation showing the data in $M$ views. The data matrix $\boldsymbol{X}^{(m)}$ could either be a feature matrix or a similarity matrix (with the link from $\boldsymbol{U}^{(m)}$ to $\boldsymbol{X}^{(m)}$ appropriately defined). Right: for $M = 3$ views, a structured-sparsity based decomposition of the view-specific latent matrices to learn shared and view-specific latent factors. First two factors are present in all the views (nonzero first two rows in each $\boldsymbol{W}^{(m)}$) while others are present only in some views. The matrix $\boldsymbol{V}$ is the global latent representation of the data.

We assume each row of $\boldsymbol{V} \in \mathbb{R}^{N \times R}$ drawn as $\boldsymbol{V}_i \sim \mathcal{N}or(\boldsymbol{0}, \boldsymbol{I})$. We use group-wise automatic relevance determination [19] as the sparsity inducing prior on $\{\boldsymbol{W}^{(m)}\}_{m=1}^M$, which also helps in inferring $R$ by shrinking the unnecessary rows in $\boldsymbol{W}$ to close to zero. Each row of $\boldsymbol{W}^{(m)}$ is assumed to be drawn as $\boldsymbol{W}_r^{(m)} \sim \mathcal{N}or(0, \alpha_{mr}^{-1}\boldsymbol{I}), r = 1, \ldots, R$, where $\alpha_{mr} \sim \mathcal{G}am(a_\alpha, b_\alpha)$ and choosing $a_\alpha, b_\alpha \to 0$, we have Jeffreys prior $p(\alpha_{mr}) \propto 1/\alpha_{mr}$, favoring strong sparsity. We can identify the factor activeness in each view from the precision hyperparameter $\alpha_{mr}$: small $\alpha_{mr}$ (large variance) indicates activeness of factor $r$ in view $m$. Let $\boldsymbol{B}$ be a $(M \times R)$-binary matrix indicating the active view vs factor associations, then $\boldsymbol{B}_{mr} = 1$ if $\alpha_{mr}^{-1} > \epsilon$, for some small $\epsilon$ (e.g., 0.01). The correlation between views $m$ and $m'$ can also be computed as $(\tilde{\boldsymbol{W}}^{(m)})^\top \tilde{\boldsymbol{W}}^{(m')} / (\boldsymbol{S}^{(m)}\boldsymbol{S}^{(m')})$ where $\tilde{\boldsymbol{W}}_r^{(m)} = \sum_{j=1}^{K_m} (\boldsymbol{W}_{rj}^{(m)})^2, r = 1, \ldots, R$ and $\boldsymbol{S}^{(m)} = \sqrt{(\tilde{\boldsymbol{W}}^{(m)})^\top \tilde{\boldsymbol{W}}^{(m')}}$.

**Identifiability via Rotation:** Factor analysis models are known to have identifiability issues due to the fact that $\boldsymbol{V}\boldsymbol{W}^{(m)} = \boldsymbol{V}\boldsymbol{Q}\boldsymbol{Q}^{-1}\boldsymbol{W}^{(m)}$, for arbitrary rotation $\boldsymbol{Q}$ [19]. We explicitly optimize w.r.t. $\boldsymbol{Q}$ to maintain identifiability in the model, and achieve faster convergence during inference.
**Adaptation for Multiview Classification:** In multiview classification, the training data consist of $N$ objects, each having $M$ feature and/or similarity based views. As earlier, we assume that the data are given as a collection of (potentially incomplete) feature and/or similarity matrices $\{\boldsymbol{X}^{(m)}\}_{m=1}^M$. Each object also has a label $y_i \in \{1, \ldots, C\}$, $i = 1, \ldots, N$, and the goal is to learn a classifier that predicts the labels for test objects where each test object has representation in $M$ views (or a subset of the views). The classification adaptation of MLFS is based on a multinomial probit model [4] on the *global* latent factors $\boldsymbol{V} = [\boldsymbol{V}_1; \ldots; \boldsymbol{V}_N]$ where $\boldsymbol{V}_i \in \mathbb{R}^{1 \times R}$, which can be summarized as: $y_i = \arg\max_c \{z_{ic}\}$, where $c = 1, \ldots, C$; $z_{ic} \sim \mathcal{N}or(\boldsymbol{V}_i\boldsymbol{\beta}_c, 1)$; $\boldsymbol{\beta}_c \sim \mathcal{N}or(\boldsymbol{0}, \rho^{-1}\boldsymbol{I})$, where $\boldsymbol{\beta}_c \in \mathbb{R}^{R \times 1}$. Under this adaptation, we learn both $\boldsymbol{V}$ and $\beta_c$ *jointly*, instead of in two separate steps.

## 3 Related Work

Most existing methods for learning from multiview data, such as [19, 20, 11] (with [20] especially focused on analyzing heterogeneous clininal data) usually either require all the views to be of the same type (e.g., all views are feature-based or all views are similarity-based), or are designed to solve specific problems on multiview data (e.g., classification or clustering or matrix completion). The idea of learning shared and view-specific latent factors for multiview data has been used in some other previous works [8, 19]. These methods however do not generalize to other feature types (e.g., ordinal/binary) or similarity matrices, and to classification/clustering problems. Another recent method [9], based on the idea of collective matrix factorization [17], jointly performs factorization of multiple matrices with each denoting a similarity matrix defined over two (from a collection of several) sets of objects (both sets can be the same). However, due to its specific construction, this method can only model a *single* similarity matrix over the objects of a given set (unlike our method which allows modeling multiple similarity matrices over the same set of objects), does not explicitly model ordinal data, does not generalize to classification/clustering, and uses a considerably different inference procedure (*batch* MAP estimation) than our proposed framework.

## 4 Experiments

We apply our proposed framework on analyzing a heterogeneous multiview data set collected from 637 college students. The data consist of 23 ordinal-valued response matrices from self-report questionnaires, concerning various behavioral/psychological aspects; one real-valued feature matrix from fMRI data having four features: threat-related (left/right) amygdala reactivity and reward-related (left/right) ventral striatum (VS) reactivity [14]; and four similarity matrices, obtained from SNP measurements of three biological systems (norepinephrine (NE), dopamine (DA) and serotonin (5-HT)) [7, 13], and a personality ratings dataset provided by *informants* (e.g., parents, sibling or friends) [18]. For the SNP data (A,C,G,T nucleotides), the similarity matrices are based on the genome-wide average proportion of alleles shared identical-by-state (IBS) [12]. For the informant reports (on 94 questions), the similarities are based on computing the averaged informants' ratings for each student and then using a similarity measure proposed in [3]. There are also binary labels associated with diagnosis of psychopathological disorders. We focus on two broadband behavioral disorders: *Internalizing* (anxious and depression symptoms) and *Externalizing* (aggressive, delinquent and hyperactive symptoms as well as substance use disorders) [10]. We apply our MLFS framework on this data to: $(i)$ interpret common/view-specific latent factors as well as view-correlations, $(ii)$ do multiview classification to predict psychopathological conditions, $(iii)$ predict missing data (e.g., question answers and fMRI response) leveraging information from multiple views. We perform analysis considering $K_m = 20$ (for similarity-based views), $R = 30$ latent factors, and prior hyperparameters $a_\alpha = b_\alpha = 0.01$.

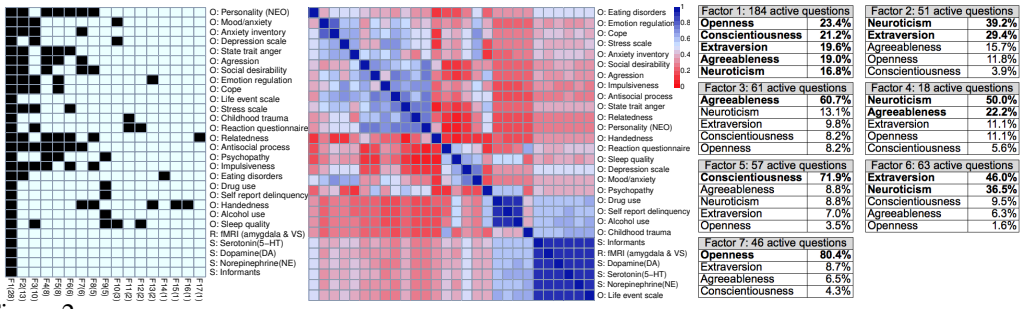


Figure 2: Left: Active views for each factor. Row labels indicate the type of view: ordinal (O), real (R) and similarity (S); column indexes factors (Number of active views in parenthesis). Middle: Inferred view-correlation matrix. Right: Type of questions associated with each one of the 7 factors for the NEO questionnaire (first row in the left panel), based on the factor loading matrix of NEO.

**Common/view-specific factors and view-correlations:** For our first task, we are interested in *understanding* the data by $(i)$ identifying latent personality traits (factors) present in the students, and $(ii)$ inferring the view-correlations. Our model can help distinguish between common and view-specific factors by looking at the view-factor association matrix $\boldsymbol{B}$ (Section 2). Figure 2 (left panel) shows the inferred view-factor associations for this data. We only show 17 factors which have at least one active view. Note that the first one represents the common factor (present in all the views), whereas the last 4 factors have only one active view (structured noise). Figure 2 (middle panel) shows the view-correlation matrix inferred from $\boldsymbol{W}^{(m)}$, computed as described in Section 2. As the figure shows, our model (seemingly) correctly discovers views that have high pairwise correlations, such as questionnaires on drug-use, self-report delinquency and alcohol-use. Further insights can

Table 1: AUC scores on the prediction of *internalizing and externalizing* disorders.

| | **MLFS (all)** | **MLFS (ordinal)** | **MLFS (real+sim.)** | **MLFS (concat.)** | **BEMKL**[5] |
|---|---|---|---|---|---|
| **Intern.** | $\mathbf{0.754 \pm 0.032}$ | $0.720 \pm 0.026$ | $0.546 \pm 0.031$ | $0.713 \pm 0.027$ | $0.686 \pm 0.037$ |
| **Extern.** | $\mathbf{0.862 \pm 0.019}$ | $0.770 \pm 0.027$ | $0.606 \pm 0.024$ | $0.747 \pm 0.034$ | $0.855 \pm 0.015$ |

be obtained by interpreting the factor loadings $\boldsymbol{W}^{(m)}$ (for which the rows correspond to factors and columns to questions). The NEO questionnaire (240 questions) is of particular interest in psychology to measure the five broad domains of personality (openness, conscientiousness, extraversion, agreeableness, and neuroticism). Figure 2 (right panel) shows, for the 7 factors active in NEO, the percentage of questions associated with every domain of personality. It is insightful to observe that the first factor includes, in an equitable manner, questions related with the five domains, whereas for the other factors, questions related with one or two domains of the personality are dominant.

**Predicting psychopathological disorders:** Our next task involves predicting each of the two types of psychopathological disorders (Internalizing and Externalizing; each is a binary classification task). To do so, we first split the data at random into training (50%) and testing (50%) sets. The training set is used to fit MLFS in four different settings: (1) MLFS with all the views, (2) MLFS with ordinal views (questionnaires), (3) MLFS with real and similarity based views (fMRI, SNP and informants), (4) MLFS concatenating the ordinal views into a *single* matrix. We consider Bayesian Efficient Multiple Kernel Learning (BEMKL) [5] as a baseline for this experiment. For this baseline, we transformed the ordinal and real-valued feature based views to kernel matrices. Each experiment repeated 10 times with different splits of training and test data. Since the labels are highly imbalanced (very few 1s), to assess the prediction performance, we compute the average of the area under ROC curve (AUC). Table 1 shows the mean AUC, bold numbers indicate the best performance. The MLFS model, which considers all the heterogeneous views, yields the overall best performance.

**Predicting ordinal responses and fMRI:** We first consider the task of ordinal matrix completion (questionnaires). We hide $(20\%, 30\%, \ldots, 90\%)$ data in each ordinal view and predict the missing data using the following methods: (1) MLFS with all the views, (2) MLFS with only ordinal views, concatenated as a single matrix, and (3) *sparse factor probit model* (SFPM) proposed in [6]. Figure 3 (top) shows the average mean absolute error (MAE) over 10 runs. The smallest MAE achieved by MLFS with all views demonstrates the benefit of integrating information from both the features and similarity based views with the group sparse factor loading matrices. Our next experiment is on predicting fMRI responses leveraging other views. For this task, we hide fMRI data from 30% of the subjects. For this group, we only assume access to the ordinal- and similarity-based views. We compare with two baselines: (1) a linear regression model (LRM) where the covariates are the ordinal responses and the similarity-based views (decomposed using SVD); (2) a *sparse* factor regression model (SFRM) [2] with same covariates as before. Figure 3 (bottom) shows the mean square error (MSE) averaged over 10 runs. Here again, MLFS outperforms the other baselines, showing the benefits of a principled generative model for the data. The Supplementary Material contains additional comparisons, including a plot for predicted vs. ground-truth of missing fMRI responses.

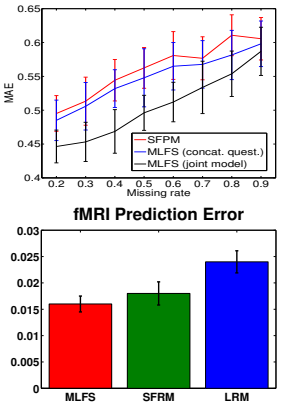

Figure 3: Predicting ordinal responses and fMRI

## 5 Conclusion

We presented a probabilistic, Bayesian framework for learning from heterogeneous multiview data consisting of diverse feature-based (ordinal, binary, real) and/or similarity-based views, with each view potentially having a significant amount of missing data. In addition to uncovering the hidden factors in multiview data, our framework allows natural adaptations for solving problems such as multiview matrix completion and multiview classification. We applied our framework for analyzing a real-world multiview clinical data, integrating the diverse data sources, uncovering latent traits of individuals, learning a classifier for psychopathological conditions, and predicting missing data for views where data acquisition may be expensive (e.g., fMRI) leveraging data from views that are inexpensive to acquire (e.g., questionnaire responses). The framework can also be easily extended for multiview clustering by replacing the multinomial probit classification model by a Gaussian mixture model over the latent factors. Finally, in the proposed framework, it is also possible to employ nonparametric Bayesian priors [15, 1] to infer the number of latent factors from data.

## References


[1] A. Bhattacharya and D. B. Dunson. Sparse bayesian infinite factor models. *Biometrika*, 98(2):291–306, 2011.

[2] C. Carvalho, J. Chang, J. Lucas, J. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling: Applications in gene expression genomics. *JASA*, 2008.

[3] A. Daemen and B. De Moor. Development of a kernel function for clinical data. In *Conf Proc IEEE Eng Med Biol Soc.*, pages 5913–5917, 2009.

[4] M. Girolami and S. Rogers. Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors. *Neural Computation*, 18, 2006.

[5] M. Gönen. Bayesian Efficient Multiple Kernel Learning. In *ICML*, 2012.

[6] P. R. Hahn, C. M. Carvalho, and J. G. Scott. A sparse factor-analytic probit model for congressional voting patterns. *Journal of the Royal Statistical Society: Series C*, 2012.

[7] A. Hariri, V. Mattay, A. Tessitore, B. Kolachana, F. Fera, D. Goldman, M. Egan, and D. Weinberger. Serotonin transporter genetic variation and the response of the human amygdala. *Science*, 297(5580):400–403, 2002.

[8] Y. Jia, M. Salzmann, and T. Darrell. Factorized Latent Spaces with Structured Sparsity. In *NIPS*, 2010.

[9] A. Klami, G. Bouchard, and A. Tripathi. Group-sparse Embeddings in Collective Matrix Factorization. *arXiv preprint arXiv:1312.5921*, 2013.

[10] R. Krueger and K. Markon. Understanding psychopathology: Melding behavior genetics, personality, and quantitative psychology to develop an empirically based model. *Current Directions in Psychological Science*, 15:113–117, 2006.

[11] A. Kumar, P. Rai, and H. Daumé III. Co-regularized Multi-view Spectral Clustering. In *NIPS*, 2011.

[12] D. J. Lawson and D. Falush. Population identification using genetic data. *Annu. Rev. Genomics Hum. Genet.*, 13:337–361, 2012.

[13] Y. Nikolova, R. Ferrell, S. Manuck, and A. Hariri. Multilocus genetic profile for dopamine signaling predicts ventral striatum reactivity. *Neuropsychopharmacology*, 36:1940–1947, 2011.

[14] Y. Nikolova and A. R. Hariri. Neural responses to threat and reward interact to predict stress-related problem drinking: A novel protective role of the amygdala. *Biology of Mood & Anxiety Disorders*, 2, 2012.

[15] P. Rai and H. Daume. The infinite hierarchical factor regression model. In *NIPS*, 2008.

[16] E. Salazar, R. Bogdan, A. Gorka, A. Hariri, and L. Carin. Exploring the Mind: Integrating Questionnaires and fMRI. In *ICML*, 2013.

[17] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *KDD*, 2008.

[18] S. Vazire. Informant reports: A cheap, fast, and easy method for personality assessment. *Journal of Research in Personality*, 40(5):472 – 481, 2006.

[19] S. Virtanen, A. Klami, S. A. Khan, and S. Kaski. Bayesian Group Factor Analysis. In *AISTATS*, 2012.

[20] S. Zhe, Z. Xu, Y. Qi, and P. Yu. Joint Association Discovery and Diagnosis of Alzheimer's Disease by Supervised Heterogeneous Multiview Learning. In *Pacific Symposium on Biocomputing*, volume 19, 2014.