Integrating Features and Similarities: Flexible Models for Heterogeneous Multiview 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

Abstract

We present a probabilistic framework for learning with *heterogeneous* multiview data where some views are given as ordinal, binary, or real-valued *feature matrices*, and some views as *similarity matrices*. Our framework has the following distinguishing aspects: (i) a unified latent factor model for integrating information from diverse feature (ordinal, binary, real) and similarity based views, and predicting the missing data in each view, leveraging view correlations; (ii) seamless adaptation to binary/multiclass classification where data consists of multiple feature and/or similarity-based views; and (iii) an efficient, variational inference algorithm which is especially flexible in modeling the views with ordinal-valued data (by *learning* the cutpoints for the ordinal data), and extends naturally to streaming data settings. Our framework subsumes methods such as multiview learning and multiple kernel learning as special cases. We demonstrate the effectiveness of our framework on several real-world and benchmarks datasets.

Introduction

Many data analysis problems involve heterogeneous data with multiple representations or *views*. We consider a general problem setting, where data in some views may be given as a *feature matrix* (which, in turn, may be ordinal, binary, or real-valued), while in other views as a *kernel or similarity matrix*. Each view may also have a significant amount of missing data. Such a problem setting is frequently encountered in diverse areas, ranging from cognitive neuroscience (Salazar et al. 2013) to recommender systems (Zhang, Cao, and Yeung 2010; Pan et al. 2011; Shi, Larson, and Hanjalic 2014). Consider a problem from cognitive neuroscience (Salazar et al. 2013), where data collected from a set of people may include ordinal-valued response matrices on multiple questionnaires, real-valued feature matrices consisting of fMRI/EEG data, and one or more similarity matrices computed using single-nucleotide polymorphism (SNP) measurements. There could also be missing observations in each view. The eventual goal could be to integrate these diverse views to learn the *latent* traits (factors) of people, or learn a classifier for predicting certain psychopathological conditions in people, or to predict the missing data in one or more views (e.g., predicting missing fMRI data, leveraging other views). Likewise, in a multi-domain recommender system (Zhang, Cao, and Yeung 2010; Pan et al. 2011; Shi, Larson, and Hanjalic 2014), for a set of users, we have multiple, partially observed ordinal-valued rating matrices for domains such as movies, books, and electronic appliances, along with a binary-valued matrix of click behavior on movies, and user-user similarities. The goal here could be to predict the missing ratings in each domain's rating matrix, leveraging information in all the sources.

When the eventual goal is only doing classification or clustering on such multiview data, one direct procedure is to represent each view (feature/kernel based) as a kernel matrix and apply multiple kernel learning methods (Gönen and Alpaydın 2011; Kumar, Rai, and Daumé III 2011). However, such an approach may be inappropriate when data is missing in one or more views, resulting in kernel matrices with missing entries, for which most of these methods are unsuitable. Moreover, these methods lack a proper generative model for each view, and therefore cannot be used for the task of *understanding* the data (e.g., learning latent factors), modeling specific feature types (e.g., ordinal), or predicting the missing data in each view (multiview matrix completion).

We present a probabilistic framework for modeling such heterogeneous multiview data with potentially missing data in each view. We then show a seamless adaptation of this framework for binary/multiclass classification for data having multiple feature- and/or similarity-based views. Our framework learns a *view-specific* latent matrix for each feature/similarity-based view, and combines these latent matrices via a set of structured-sparsity driven factor analysis models (Virtanen et al. 2012) to learn a *global* low-dimensional representation of the data. The view-specific latent matrices can be used for matrix completion for each view, whereas the global representation can be combined with specific objectives to solve problems such as classification or clustering of the multiview data. Our framework also consists of an efficient, variational inference algorithm, which 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 data, such as in recommender systems.

A Generative Framework for Heterogeneous Multiview Data

We first describe our basic framework for **M**ultiview **L**earning with **F**eatures and **S**imilarities (abbreviated henceforth as MLFS), for modeling heterogeneous multiview data, where the views may be in the form of ordinal/binary/real-valued feature matrices and/or real-valued similarity matrices. Our framework enables one to integrate the data from all the views to learn latent factors underlying the data, predict missing data in each view, and infer view-correlations. We then show how MLFS can be adapted for binary/multiclass classification problems.

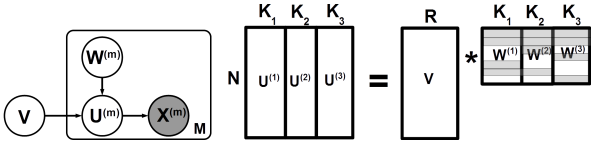
We assume the data consist of N objects having a total of M feature-based and/or similarity-based views. Of the $M = M_1 + M_2 + M_3$ views, the first M_1 are assumed to be ordinal feature matrices $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(M_1)}$ (binary feature matrix is a special case), the next 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 described in the next Section.

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 of the entity under test (in addition to capturing aspects that are present in all views), we wish to impose this structure in the learned global latent factor matrix $\boldsymbol{V}$ and the associated factor loading matrices $\boldsymbol{W} = \{\boldsymbol{W}^{(m)}\}_{m=1}^M$. Note that each column (resp., row) of $\boldsymbol{V}$ (resp., $\boldsymbol{W}$) corresponds to a global latent factor. We impose a structured-sparsity prior in the factor loading matrices $\{\boldsymbol{W}^{(m)}\}_{m=1}^M$ 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.



(a) (b)

Figure 1: (a) 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). (b) 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 (Virtanen et al. 2012) 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 α_{mr}: small α_{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 ϵ (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 $\boldsymbol{Q}$ (Virtanen et al. 2012). We explicitly optimize w.r.t. $\boldsymbol{Q}$ to maintain identifiability in the model, and achieve faster convergence during inference.

Adaptation for Multiview Classification

This general framework for MLFS can be applied for multiview factor analysis and matrix completion problems when data consists of multiple feature-based (ordinal/binary/real) and/or similarity-based views. Our framework is more general than other prior works for these problems, that assume all views having the same *feature-based* representation (Virtanen et al. 2012; Zhang, Cao, and Yeung 2010). We now show how MLFS can be adapted for other problems such as 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 (Girolami and Rogers 2006) 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 β_c *jointly*, instead of in two separate steps.

A particular advantage of our framework for classification is that, in addition to handling views having potentially different representations, it allows objects to be missing in one or more views. The existing multiview or multiple kernel learning (Yu et al. 2011; Gönen and Alpaydın 2011) methods require the views to be transformed into the same representation and/or cannot easily handle missing data. A similar adaptation can be used to perform clustering instead of multiclass classification by replacing the multinomial probit classification by a Gaussian mixture model.

Inference

Since exact inference is intractable, we use variational Bayesian EM (Beal 2003) to perform approximate inference. We will infer the variational distribution for the latent variables, collectively referred to as $\boldsymbol{\Theta}$, and consisting of $\{\{\boldsymbol{U}^{(m)}, \boldsymbol{W}^{(m)}, \boldsymbol{\alpha}_m, \gamma_m\}_{m=1}^{M}, \boldsymbol{V}\}$, along with $\{\boldsymbol{\beta}_c, \boldsymbol{z}_c\}_{c=1}^{C}$ for classification. For the cutpoints $\boldsymbol{G} = \{\boldsymbol{g}^m\}_{m=1}^{M_1}$ and the rotation matrix $\boldsymbol{Q}$, we will seek a point estimate. As will be shown, our inference algorithm also works in the streaming setting (Broderick et al. 2013) where each data point is seen only once. Sometimes, for brevity, we will use $\{\boldsymbol{U}, \boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\gamma}\}$ for $\{\boldsymbol{U}^{(m)}, \boldsymbol{W}^{(m)}, \boldsymbol{\alpha}_m, \gamma_m\}_{m=1}^{M}$, $\{\boldsymbol{\beta}, \boldsymbol{z}\}$ for $\{\{\boldsymbol{\beta}_c\}_{c=1}^{C}, \boldsymbol{z}\}$, and $\{\boldsymbol{\mu}, \boldsymbol{z}\}$ for $\{\{\boldsymbol{\mu}_j\}_{j=1}^{J}, \boldsymbol{z}\}$, respectively. The data from all views will be collectively referred to as $\boldsymbol{\mathcal{X}}$ which, in the classification case, also consists of the labels. Due to the lack of space, we only provide brief descriptions of the key aspects of our inference algorithm, leaving further details in the Supplementary Material.

We approximate the true posterior $p(\boldsymbol{\Theta}|\boldsymbol{\mathcal{X}}, \boldsymbol{G}, \boldsymbol{Q}))$ by its mean-field approximation:

$$q(\boldsymbol{\Theta}) = \prod_{m=M_1+M_2+1}^{M} \prod_{i=1}^{N} q(\boldsymbol{U}_i^{(m)}) \prod_{i=1}^{N} q(\boldsymbol{V}_i)$$

$$\prod_{m=1}^{M} \prod_{r=1}^{R} q(\boldsymbol{W}_r^{(m)}) \prod_{m=1}^{M} \prod_{r=1}^{R} q(\alpha_{mr}) \prod_{m=1}^{M} q(\gamma_m) \quad (1)$$

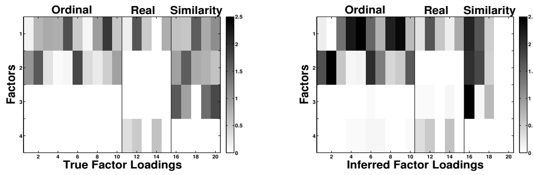
Thus, we minimize the KL-divergence $KL(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}|\boldsymbol{\mathcal{X}}, \boldsymbol{G}, \boldsymbol{Q}))$, equivalent to maximizing the evidence lower bound (ELBO) given by $\mathcal{L}(q(\boldsymbol{\Theta}), \boldsymbol{G}, \boldsymbol{Q}) = \mathbb{E}_{q(\boldsymbol{\Theta})}[\log p(\boldsymbol{\mathcal{X}}, \boldsymbol{\Theta}|\boldsymbol{G}, \boldsymbol{Q}) - \log(q(\boldsymbol{\Theta}))]$. With further approximation for the ordinal-valued views using a Taylor-series expansion, we can efficiently update variational parameters for $q(\boldsymbol{\Theta})$. Note that in (1), the terms $q(\boldsymbol{U}_i^{(m)})$ appear only for the similarity-based views because for feature-based view, $\boldsymbol{U}_i^{(m)}$ are integrated out in the variational lower bound (details of derivation are provided in Supplementary Material). We maximize the variational lower bound by iterating between variational E-step: $\max_{q(\boldsymbol{\Theta})} \mathcal{L}(q(\boldsymbol{\Theta}), \boldsymbol{G}, \boldsymbol{Q})$, and M-step: $\max_{\boldsymbol{G},\boldsymbol{Q}} \mathcal{L}(q(\boldsymbol{\Theta}), \boldsymbol{G}, \boldsymbol{Q})$. In this section, we summarize the variational updates for $\boldsymbol{U}$, $\boldsymbol{W}$, $\boldsymbol{V}$, the cutpoints $\boldsymbol{G}$, the rotation matrix $\boldsymbol{Q}$, and the extension to the streaming setting.

Update $\boldsymbol{U}_i^{(m)}$: For similarity-based views, the variational posterior $q(\boldsymbol{U}_i^{(m)}) = \mathcal{N}or(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u)$, where $\boldsymbol{\Sigma}_u = (\langle\gamma_m\rangle \boldsymbol{I}_{K_m} + \sum_{j\neq i}\langle\tau_m\rangle\langle\boldsymbol{U}_j^{(m)\top}\boldsymbol{U}_j^{(m)}\rangle)^{-1}$ and $\boldsymbol{\mu}_u = (\langle\gamma_m\rangle\langle\boldsymbol{V}_i\rangle\langle\boldsymbol{W}^{(m)}\rangle + \langle\tau_m\rangle(\sum_{j>i}\boldsymbol{X}_{ij}^{(m)}\langle\boldsymbol{U}_j^{(m)}\rangle + \sum_{j<i}\boldsymbol{X}_{ji}^{(m)}\langle\boldsymbol{U}_j^{(m)}\rangle)\boldsymbol{\Sigma}_u$, where $\langle.\rangle$ denotes expectation w.r.t. q.

Update $\boldsymbol{V}_i$: $q(\boldsymbol{V}_i) = \mathcal{N}or(\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v)$. For the supervised multiclass classification case with $\mathcal{N}or(\boldsymbol{0}, \boldsymbol{I})$ prior on $\boldsymbol{V}_i$, $\boldsymbol{\Sigma}_v = (\boldsymbol{I} + \sum_{m=1}^{M}\langle\gamma_m\rangle\langle\boldsymbol{W}^{(m)}\boldsymbol{W}^{(m)\top}\rangle + \sum_{c=1}^{C}\langle\boldsymbol{\beta}_c\boldsymbol{\beta}_c^\top\rangle)^{-1}$ and $\boldsymbol{\mu}_v = (\sum_{m=1}^{M_1}\langle\gamma_m\rangle\sum_{j=1}^{K_m}\langle\boldsymbol{W}_{:j}^{(m)\top}\rangle(\boldsymbol{g}^m_{\boldsymbol{X}_{i,j}^{(m)}} + \boldsymbol{g}^m_{\boldsymbol{X}_{i,j}^{(m)}-1})/2 + \sum_{m=M_1+1}^{M_1+M_2}\langle\gamma_m\rangle\boldsymbol{X}_i^{(m)}\langle\boldsymbol{W}^{(m)\top}\rangle + \sum_{m=M_1+M_2+1}^{M_1+M_2+M_3}\langle\gamma_m\rangle\langle\boldsymbol{U}_i^{(m)}\rangle\langle\boldsymbol{W}^{(m)\top}\rangle + \sum_{c=1}^{C}\langle z_{ic}\rangle\langle\boldsymbol{\beta}_c^\top\rangle)\boldsymbol{\Sigma}_v$.

Update $\boldsymbol{W}^{(m)}$: The variational posterior for the j-th column of $\boldsymbol{W}^{(m)}$ ($j = 1, \ldots, K_m$), is given by $q(\boldsymbol{W}_{:j}^{(m)}) = \mathcal{N}or(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$, where $\boldsymbol{\Sigma}_w = (\langle\text{diag}(\alpha_{m1}, \alpha_{m2}, ..., \alpha_{mR})\rangle + \langle\gamma_t\rangle\sum_{i=1}^{N}\langle\boldsymbol{V}_i^\top\boldsymbol{V}_i\rangle)^{-1}$ and $\boldsymbol{\mu}_w$ depends on the type of view. For ordinal-valued feature-based view, $\boldsymbol{\mu}_w = \boldsymbol{\Sigma}_w\langle\gamma_m\rangle\sum_{i=1}^{N}\langle\boldsymbol{V}_i^\top\rangle(\boldsymbol{g}^m_{\boldsymbol{X}_{i,j}^{(m)}} + \boldsymbol{g}^m_{\boldsymbol{X}_{i,j}^{(m)}-1})/2$. For real-valued feature-based view, $\boldsymbol{\mu}_w = \boldsymbol{\Sigma}_w\langle\gamma_m\rangle\sum_{i=1}^{N}\langle\boldsymbol{V}_i^\top\rangle\boldsymbol{X}_{ij}^{(m)}$. For similarity-based view, $\boldsymbol{\mu}_w = \boldsymbol{\Sigma}_w\langle\gamma_m\rangle\sum_{i=1}^{N}\langle\boldsymbol{V}_i^\top\rangle\langle\boldsymbol{U}_{ij}^{(m)}\rangle$.

Inferring the cutpoints: We infer the cutpoints $\boldsymbol{g}^m$ for each ordinal view by optimizing the objective function $\tilde{\mathcal{L}}^m(\boldsymbol{g}^m) = \sum_{l=1}^{L_m}\tilde{\mathcal{L}}_l^m$, where $\tilde{\mathcal{L}}_l^m = N_l^m[\log(\boldsymbol{g}_l^m - \boldsymbol{g}_{l-1}^m) - \langle\gamma_m\rangle(\boldsymbol{g}_l^{m\,2} + \boldsymbol{g}_{l-1}^{m}{}^2 + \boldsymbol{g}_l^m\boldsymbol{g}_{l-1}^m)/6] + \langle\gamma_m\rangle(\boldsymbol{g}_l^m +$



(a) (b)

Figure 2: (a) and (b): true and inferred factor loading matrices $\boldsymbol{W}$ for ordinal, real, and similarity-based views on synthetic data.

$\boldsymbol{g}^m_{l-1}) \sum_{i,j:\boldsymbol{X}^m_{i,j}=l} \langle \boldsymbol{V}_i \rangle \langle \boldsymbol{W}^{(m)}_{:j} \rangle / 2$. Here N^m_l is the number of observations with value l in view m. The gradients of $\tilde{\mathcal{L}}^t_l$ are also analytically available. Moreover, the objective function is concave w.r.t. $\boldsymbol{g}^m$ – in each variational M-step, the solution $\hat{\boldsymbol{g}}^m$ given the variational distributions $q(\boldsymbol{\Theta})$ is globally optimal. It can be solved efficiently using Newton's method. At every VB iteration, optimization over $\boldsymbol{g}^m$ is guaranteed to increase the variational lower bound.

Inferring the rotation matrix $\boldsymbol{Q}$: Since rotation leaves $p(\boldsymbol{X}|\boldsymbol{\Theta})$ unchanged, optimization over $\boldsymbol{Q}$ effectively minimizes $KL(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}))$ (encourages $\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\alpha}$ to be similar to the prior). This encourages structured sparsity (imposed by prior), helps escaping from local optima, and achieves faster convergence (Virtanen et al. 2012). Figure 2 shows an experiment on a three-view synthetic data, consisting of ordinal, real, and similarity based views, each generated from a subset of four latent factors. The rotation matrix also helps in the identifiability of inferred loading matrices.

Streaming extension: In the streaming setting, given $\boldsymbol{X}_n$ (a new data point, or a batch of new data points, with each having some or all the views), we infer the latent factors $\boldsymbol{V}_n$ using the following update equations: $q(\boldsymbol{V}_n) = \mathcal{N}or(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, where $\boldsymbol{\mu}_n = (\sum_{m=1}^{M_1} \sum_j \langle \gamma_m \rangle \langle {\boldsymbol{W}^{(m)}_{:j}}^\top \rangle (\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{n,j}} + \boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{n,j}-1})/2 + \sum_{m=M_1+1}^{M_1+M_2} \langle \gamma_m \rangle \boldsymbol{X}^{(m)}_n \langle {\boldsymbol{W}^{(m)}}^\top \rangle + \sum_{m=M_1+M_2+1}^{M_1+M_2+M_3} \langle \gamma_m \rangle \langle \boldsymbol{U}^{(m)}_n \rangle \langle {\boldsymbol{W}^{(m)}}^\top \rangle) \boldsymbol{\Sigma}_n$, and $\boldsymbol{\Sigma}_n = (\boldsymbol{I} + \sum_{m=1}^{M} \langle \gamma_m \rangle \langle \boldsymbol{W}^{(m)} {\boldsymbol{W}^{(m)}}^\top \rangle)^{-1}$. The global variables $\boldsymbol{\Theta}$, such as $\boldsymbol{W}^{(m)}$, can be updated in a manner similar to (Broderick et al. 2013) as $q(\boldsymbol{\Theta}_n) \propto q(\boldsymbol{\Theta}_{n-1}) p(\boldsymbol{X}_n|\boldsymbol{\Theta}_n)$. Due to the lack of space, the experiments for the streaming setting are presented separately in the Supplementary Material.

Related Work

The existing methods for learning from multiview data, such as (Gönen and Alpaydın 2011; Virtanen et al. 2012; Zhang, Cao, and Yeung 2010; Bickel and Scheffer 2004; Yu et al. 2011; Shao, Shi, and Yu 2013; Zhe et al. 2014; Chaudhuri et al. 2009; Kumar, Rai, and Daumé III 2011), usually either require all the views to be of the same type (e.g., feature based or similarity based), or are designed to solve specific problems on multiview data (e.g., classification or clustering or matrix completion). Moreover, most of these are non-generative w.r.t. the views (Gönen and Alpaydın 2011; Chaudhuri et al. 2009; Kumar, Rai, and Daumé III 2011), lacking a principled mechanism to handle/predict missing data in one or more views. The idea of learning shared and view-specific latent factors for multiview data has been used in some other previous works (Jia, Salzmann, and Darrell 2010; Virtanen et al. 2012). 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 (Klami, Bouchard, and Tripathi 2014), based on the idea of collective matrix factorization (Singh and Gordon 2008), 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.

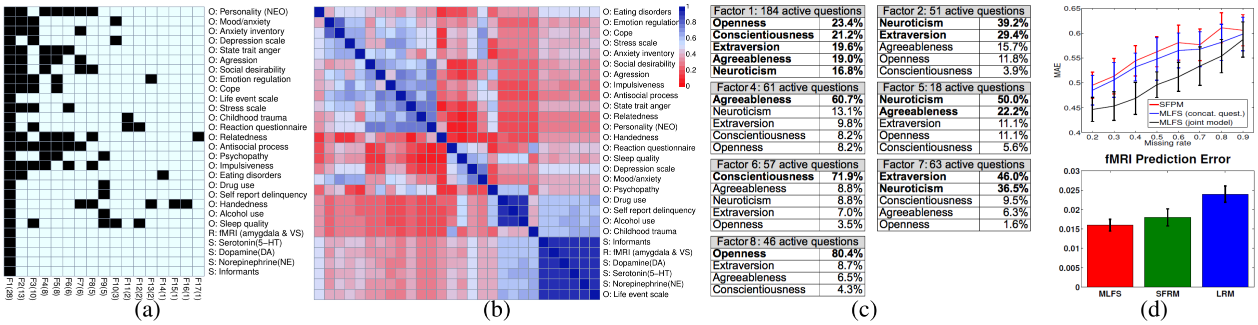
Finally, related to our contribution on inferring the cutpoints for the ordinal views, in another recent work (Hernandez-Lobato, Houlsby, and Ghahramani 2014) proposed a single-view ordinal matrix completion model with an Expectation Propagation (EP) based algorithm for learning the cutpoints. It however assumes a Gaussian prior on the cutpoints. We do not make any such assumption and *optimize* w.r.t. the cutpoints. Moreover, the optimization problem for inferring the cutpoints is concave, leading to an efficient and fast-converging inference.

Experiments

In this section, we first apply our framework for analyzing a real-world dataset from cognitive neuroscience. We then present results on benchmark datasets for recommender system matrix completion and classification, respectively.

Cognitive Neuroscience Data

This is a heterogeneous multiview data 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 (Nikolova and Hariri 2012); and four similarity matrices, obtained from SNP measurements of three biological systems (norepinephrine (NE), dopamine (DA) and serotonin (5-HT)) (Hariri et al. 2002; Nikolova et al. 2011), and a personality ratings dataset provided by *informants* (e.g., parents, sibling or friends) (Vazire 2006). 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) (Lawson and Falush 2012). 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 (Daemen and De Moor 2009). There are also binary labels associated with diagnosis of psychopathological disorders. We



Factor 1: 184 active questions	
Openness	**23.4%**
Conscientiousness	**21.2%**
Extraversion	**19.6%**
Agreeableness	**19.0%**
Neuroticism	**16.8%**

Factor 2: 51 active questions	
Neuroticism	**39.2%**
Extraversion	**29.4%**
Agreeableness	15.7%
Openness	11.8%
Conscientiousness	3.9%

Factor 4: 61 active questions	
Agreeableness	**60.7%**
Neuroticism	13.1%
Extraversion	9.8%
Conscientiousness	8.2%
Openness	8.2%

Factor 5: 18 active questions	
Neuroticism	**50.0%**
Agreeableness	**22.2%**
Extraversion	11.1%
Openness	11.1%
Conscientiousness	5.6%

Factor 6: 57 active questions	
Conscientiousness	**71.9%**
Agreeableness	8.8%
Neuroticism	8.8%
Extraversion	7.0%
Openness	3.5%

Factor 7: 63 active questions	
Extraversion	**46.0%**
Neuroticism	**36.5%**
Conscientiousness	9.5%
Agreeableness	6.3%
Openness	1.6%

Factor 8: 46 active questions	
Openness	**80.4%**
Extraversion	8.7%
Agreeableness	6.5%
Conscientiousness	4.3%

Figure 3: (a) 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). (b) Inferred view-correlation matrix. (c) 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. (d) Predicting ordinal responses and predicting fMRI.

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

	MLFS (all)	MLFS (ordinal)	MLFS (real+sim.)	MLFS (concat.)	BEMKL
Intern.	**0.754 ± 0.032**	0.720 ± 0.026	0.546 ± 0.031	0.713 ± 0.027	0.686 ± 0.037
Extern.	**0.862 ± 0.019**	0.770 ± 0.027	0.606 ± 0.024	0.747 ± 0.034	0.855 ± 0.015

focus on two broadband behavioral disorders: *Internalizing* (anxious and depression symptoms) and *Externalizing* (aggressive, delinquent and hyperactive symptoms as well as substance use disorders) (Krueger and Markon 2006). 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 $a_\alpha = b_\alpha = 0.01$.

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 $\boldsymbol{B}$ (model section). Figure 3(a) 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 3(b) shows the view-correlation matrix inferred from $\boldsymbol{W}^{(m)}$, computed as described in the model section. 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 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 3(c) 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 predicts 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) (Gönen 2012) as a baseline for this experiment. For this baseline, we transformed the ordinal and real-valued feature based views to kernel matrices. Each experiment is 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 with standard deviation, 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 (Hahn, Carvalho, and Scott 2012). Top plot in Figure 3 (d) 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

Table 2: Benchmark datasets: Ordinal matrix completion leveraging the similarity based view.

	Epinion		**Ciao**	
	MAE	**Exact Match**	**MAE**	**Exact Match**
Ordinal only	0.8700 (±0.0079)	0.3871 (±0.0056)	1.0423 (±0.0162)	0.3039 (±0.0068)
KPMF	1.0664 (±0.0204)	0.2715 (± 0.0212)	1.1477(±0.0242)	0.2788 (±0.0140)
MLFS	**0.8470 (±0.0050)**	**0.4060 (±0.0102)**	**0.9826 (±0.0133)**	**0.3261 (±0.0070)**

Table 3: Benchmark datasets: Accuracies on multiple similarity matrix based classification.

	UCI Handwritten Digits (10 classes)		**Protein Fold (27 classes)**	
	No missing	**50% missing**	**No missing**	**50% missing**
Concatenation	93.47% (±1.40%)	92.02% (±2.03%)	50.46% (± 2.96%)	45.93% (± 3.59%)
BEMKL	94.94% (±0.84%)	88.59% (±2.76%)	**53.70% (±2.88%)**	47.77% (±3.01%)
MLFS	**95.14% (±0.85%)**	**93.61% (±1.16%)**	51.11% (±2.02%)	**48.27% (±2.48%)**

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) (Carvalho et al. 2008) with same covariates as before. Bottom plot in Figure 3 (d) 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.

Matrix Completion for Recommender Systems

For this task, we consider two benchmark datasets[1], Epinion and Ciao, both having two views: ordinal rating matrix (range 1-5) and similarity matrix. The goal in this experiment is to complete the partially observed rating matrix, leveraging similarity based view. Note that multiview matrix completion methods such as (Zhang, Cao, and Yeung 2010) cannot be applied for this task because these require all the views to be of the same type (e.g., ordinal matrix). The Epinion dataset we use consists of a 1000×1500 ordinal user-movie rating matrix ($\sim$ 2% observed entries). The Ciao dataset we use consists of product 1000×500 ordinal user-DVDs rating matrix ($\sim$ 1% observed entries). In addition, for each dataset, we are given a network over the users which is converted into a 1000×1000 similarity matrix (computed based on the number of common trusted users for each pair of users). We compare our method with two baselines: (i) Ordinal only: uses only the ordinal view, (ii) Kernelized Probabilistic Matrix Factorization (KPMF) (Zhou et al. 2012): allows using a similarity matrix to assist a matrix completion task (it however treats the ratings as real-valued; we round it when comparing the exact match). We run 10 different splits with 50% of the observed ratings as training set and the remaining 50% ratings as test set and report the averaged results. As Table 2 shows, our method outperforms both baselines in terms of the completion accuracy (Mean-Absolute Error and Exact Match).

Multiview/Multiple Kernel Classification

Our next experiment is on the task of multiple kernel classification on benchmark datasets. The multinomial probit adaptation of MLFS, with all similarity-based views, naturally applies for this problem. For this experiment, we choose two benchmark datasets: UCI Handwritten Digits (Kumar, Rai, and Daumé III 2011) and Protein Fold Prediction (Gönen 2012). The Digits data consists of 2000 digits (10 classes), with each having 6 type of feature representations. We construct 6 kernel matrices for this data in the same manner as (Kumar, Rai, and Daumé III 2011). We split the data into 100 digits for training and 1900 digits for test. The Protein data consists of 624 protein samples (27 classes), each having 12 views. We construct 12 kernel matrices for this data in the same manner as (Gönen 2012). For Protein data, we split the data equally into training and test sets. For both Digits and Protein data experiments, for each training/test split (10 runs), we try two settings: no missing and 50% missing observations in each view. We compare with two baselines: (i) Concatenation: performs SVD on each view's similarity matrix, concatenates all of the resulting matrices, and learns a multiclass probit model, and (ii) Bayesian Efficient Multiple Kernel Learning (BEMKL) (Gönen 2012), which is a state-of-the-art multiple kernel learning method. The results are shown in Table 3. For the missing data setting, we use zero-imputation for the baseline methods (our method does not require imputation). As shown in the table, our method yields better test set classification accuracies as compared to the other baselines. For Protein data, although BEMKL performs better for the fully observed case, our method is better when the data in each view is significantly missing.

Conclusion

We presented a probabilistic, Bayesian framework for learning from heterogeneous multiview data consisting of diverse feature-based (ordinal, binary, real) and similarity-based views. In addition to learning the latent factors and view correlations in multiview data, our framework allows solving various other problems involving multiview data, such as matrix completion and classification. Our contribution on learning the cutpoints for ordinal data is useful in its own right (e.g., applications in recommender systems). The streaming extension shows the feasibility of posing our framework in online learning and active learning, left as future work. Our work can also be extended for multiview clustering when the data consists a mixture of feature- (real, binary, ordinal, etc.) and similarity-based views.

Acknowledgments

The research reported here was funded in part by ARO, DARPA, DOE, NGA and ONR.

[1] http://www.public.asu.edu/ jtang20/datasetcode/truststudy.htm/

References

[Beal 2003] Beal, M. J. 2003. *Variational algorithms for approximate Bayesian inference*. Ph.D. Dissertation, University of London.

[Bickel and Scheffer 2004] Bickel, S., and Scheffer, T. 2004. Multi-View Clustering. In *ICDM*.

[Broderick et al. 2013] Broderick, T.; Boyd, N.; Wibisono, A.; Wilson, A. C.; and Jordan, M. 2013. Streaming Variational Bayes. In *NIPS*.

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

[Chaudhuri et al. 2009] Chaudhuri, K.; Kakade, S. M.; Livescu, K.; and Sridharan, K. 2009. Multi-view Clustering via Canonical Correlation Analysis. In *ICML*.

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

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

[Gönen and Alpaydın 2011] Gönen, M., and Alpaydın, E. 2011. Multiple Kernel Learning Algorithms. *JMLR*.

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

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

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

[Hernandez-Lobato, Houlsby, and Ghahramani 2014] Hernandez-Lobato, J. M.; Houlsby, N.; and Ghahramani, Z. 2014. Probabilistic matrix factorization with non-random missing data. In *ICML*.

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

[Klami, Bouchard, and Tripathi 2014] Klami, A.; Bouchard, G.; and Tripathi, A. 2014. Group-sparse Embeddings in Collective Matrix Factorization. In *ICLR*.

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

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

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

[Nikolova and Hariri 2012] Nikolova, Y., and Hariri, A. R. 2012. 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.

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

[Pan et al. 2011] Pan, W.; Liu, N. N.; Xiang, E. W.; and Yang, Q. 2011. Transfer Learning to Predict Missing Ratings via Heterogeneous User Feedbacks. In *IJCAI*.

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

[Shao, Shi, and Yu 2013] Shao, W.; Shi, X.; and Yu, P. 2013. Clustering on Multiple Incomplete Datasets via Collective Kernel Learning. *arXiv preprint arXiv:1310.1177*.

[Shi, Larson, and Hanjalic 2014] Shi, Y.; Larson, M.; and Hanjalic, A. 2014. Collaborative Filtering Beyond the User-Item Matrix: A Survey of the State of the Art and Future Challenges. *ACM Comput. Surv.* 47(1).

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

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

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

[Yu et al. 2011] Yu, S.; Krishnapuram, B.; Rosales, R.; and Rao, R. B. 2011. Bayesian Co-Training. *JMLR*.

[Zhang, Cao, and Yeung 2010] Zhang, Y.; Cao, B.; and Yeung, D. 2010. Multi-Domain Collaborative Filtering. In *UAI*.

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

[Zhou et al. 2012] Zhou, T.; Shan, H.; Banerjee, A.; and Sapiro, G. 2012. Kernelized Probabilistic Matrix Factorization: Exploiting Graphs and Side Information. In *SDM*.