A Bayesian Framework for Multi-Modality Analysis of Mental Health

Esther Salazar[1], Yuliya Nikolova[2], Wenzhao Lian[1], Piyush Rai[1], Adrienne L. Romer[2], Ahmad R. Hariri[2], and Lawrence Carin[1]

[1]*Electrical and Computer Engineering Department*
[2]*Department of Psychology & Neuroscience*
Duke University
Durham, NC 27708-0291

Abstract

We develop statistical methods for multi-modality assessment of mental health, based on four forms of data: (*i*) self-reported answers to a set of classical questionnaires, (*ii*) single-nucleotide polymorphism (SNP) data, (*iii*) fMRI data measured in response to visual stimuli, and (*iv*) scores for psychiatric disorders. The data were acquired from hundreds of college students. We utilize the data and model to ask a timely and novel clinical question: can one predict brain activity associated with risk for mental illness and treatment response based on knowledge of how the subject answers questionnaires, and using genetic (SNP) data? Also, in another direction: can one predict an individual's fundamental propensity for psychopathology based on observed self-report, SNP and fMRI data (separately or in combination)? The data are analyzed with a multi-modality factor model, with sparsity imposed on the factor loadings, linked to the particular type of data modality. The analysis framework encompasses a wide range of problems, such as matrix completion and clustering, leveraging information in all the data sources. We use an efficient variational inference algorithm to fit the model, which is especially flexible in dealing with ordinal-valued views (self-report answers and SNP data). The variational inference is validated with slower but rigorous sampling methods. We demonstrate the effectiveness of the model to perform accurate predictions for clinically relevant brain activity relative to baseline models, and to identify meaningful associations between data views.

Key words: Bayesian factor model; Clustering; Multi-view learning; Neuroscience

1 Introduction

1.1 Background and motivation

There is intense scientific, clinical, and public health interest in finding neural and genetic biomarkers that are early-warning signs for psychiatric diseases, as well as potential prevention targets (Cook, 2008; Singh and Rose, 2009). This interest has been bolstered by technological advances in the fields of neuroimaging and neurogenetics, as well as concerted efforts to construct large databases that afford the power to detect likely small effects of individual risk biomarkers (Nikolova et al., 2013; Schumann et al., 2010; Stein et al., 2012).

Nevertheless, significant challenges remain to progress in identifying biomarkers of actual clinical utility at the population level.

For example, growing evidence indicates that assays based on functional magnetic resonance imaging (fMRI) (Huettel et al., 2009) of how specific neural circuits process reward and threat can accurately classify individuals into risk groups and illuminate novel targets for intervention and prevention (Fakra et al., 2009; Forbes et al., 2009; Nikolova et al., 2012; Nikolova and Hariri, 2012). However, conducting fMRI assays of neural circuit function is non-trivial, requiring not only significant financial resources but also an extensive research infrastructure to administer assays properly. It is thus unlikely that fMRI will be implemented in clinical settings at a population level.

In contrast to fMRI assays of neural circuit function, common variation in our DNA sequence (i.e., genetic polymorphisms) is relatively inexpensive to assay from easily accessible peripheral tissues, such as saliva. A growing body of research from the field of neurogenetics has begun to identify specific functional genetic variants which account for small but significant inter-individual variability in neural circuit function associated with disease risk (Hariri, 2009a). As such, these genetic markers have the potential to represent proxies for neural phenotypes that are biomarkers of risk.

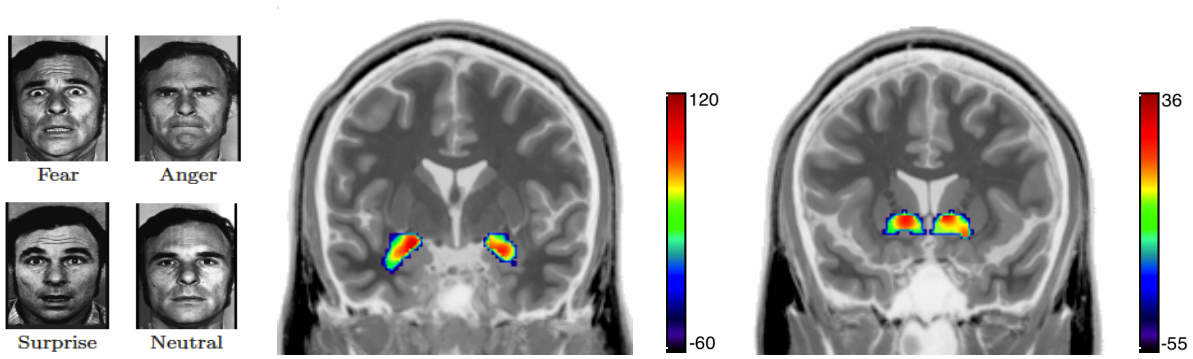
Responses to standardized pencil-and-paper or online self-report measures of dispositional traits, such as personality, represent another potential source of easily assayed proxies of neural risk phenotypes. Like genetic polymorphisms, individual differences in self-report measures correlate with small but significant variability in neural circuit function (Fakra et al., 2009; Forbes et al., 2009; Nikolova et al., 2013). Thus, genetic polymorphisms and self-reported dispositional traits, either independently or in combination, have the potential to serve as relatively easily assayed proxies of neural risk biomarkers that can be scaled in a manner than can inform and advance mental health at the population level. Despite the considerable promise of genetic and self-report measures as scalable proxies of neural risk biomarkers, existing research has largely been limited to studies employing a single or very small number of genetic or self-report variables, to account for variability in behaviorally and clinically relevant neural circuit function (Nikolova et al., 2013). Not surprisingly, these studies have accounted for only a small amount of variance in neural circuit function, often using antiquated approaches for statistical inference such as standard regression models.

1.2 Data description

In this paper we draw upon data collected through the ongoing Duke Neurogenetics Study (DNS) (Carre et al., 2013; Nikolova and Hariri, 2012). The data considered here are collected from $N = 653$ Duke University undergraduate students, age 18–22 years, with 58% female. All data were collected under guidelines stipulated by a Duke University Institutional Review Board (IRB). As a component of the DNS, each participant completed two well-characterized fMRI challenge paradigms to assess threat-related amygdala reactivity and reward-related ventral striatum (VS) reactivity.

The first challenge paradigm elicited amygdala reactivity to canonical facial expressions signaling various forms of threat in the environment (Ahs et al., 2014). Each participant was asked over repeated trials to indicate with a button press which one of two faces shown on the bottom of a screen is identical to a target face shown on top of the screen. Notably, the faces had one of the following expressions: anger, fear, surprise or neutral (see Figure 1(a) for a prototypical example). Four task blocks of expression-specific face matching were interleaved with five control blocks of matching shapes (circles and ovals). Following data preprocessing, four relevant contrasts were created reflecting the reactivity of the amygdala to each of the four facial expressions, relative to the control condition (e.g., Anger>Shapes, Fear>Shapes, Surprise>Shapes, and Neutral>Shapes). The fMRI-measured data from this paradigm was summarized in terms of 10 real matrices (for each matrix, rows representing people, columns holding values for a certain number of spatial voxels). Specifically, the first 8 real matrices are associated with amygdala reactivity for each hemisphere (left and right); two fMRI responses for each of the four contrasts.

The second fMRI paradigm, which probed reward-related ventral striatum (VS) reactivity, consisted of a number-guessing task, wherein correct guesses of a numerical value were rewarded and incorrect guesses were penalized with positive and negative feedback respectively (Hariri et al., 2006). For each subject, we have fMRI voxels values reflecting reward-related reactivity in the left and right ventral striatum (VS). Example fMRI data are shown in Figure 1(b)-(c) for amygdala and VS reactivity (left and right hemisphere) corresponding to one subject in the sample. The left and right amygdala are characterized by 167 and 194 voxels, respectively, while for the VS the left and right regions are characterized by 301 and 329 voxels, respectively. Concerning the asymmetry in the number of voxels on the left and right regions, there is asymmetry in the number of voxels that exhibit a statistically significant response to our fMRI tasks. With the amygdala, there is further asymmetry in the size of the anatomical regions of interest to begin with. The size of the voxels is con-



(a) Stimuli (b) Amygdala (c) Ventral striatum

Figure 1: Panel (a): Prototypical examples of the four facial expressions used for visual stimuli. Panels (b)-(c): Threat-related reactivity in the amygdala (b) and ventral striatum-VS (c), respectively. Single-subject values (scaled by 100) are shown onto a structural MRI template in the coronal plane.

strained by the acquisition parameters of our fMRI sequence, which strives to balance spatial resolution and coverage of the whole brain. Each fMRI voxel is $2mm \times 2mm \times 2mm$.

We targeted threat-related amygdala and reward-related VS reactivity as our behaviorally and clinically relevant fMRI measures for several reasons (Nikolova and Hariri, 2012). First, amygdala activity is critical for generating adaptive changes in behavior and physiology in response to threat in our environment, while VS activity is crucial for effecting similar changes in response to rewards. A second related feature is a generally robust reactivity of the amygdala and the VS, which is readily measured with fMRI, to threat- and reward-related visual cues, respectively. Third, fMRI measures of both amygdala and VS reactivity exhibit considerable inter-individual variability, which maps onto individual differences in many self-report measures, including those assessing personality, mood, affect, and the experience of stress (Hariri, 2009b). Finally, the amygdala and the VS are clearly important in the emergence of mental illness, particularly mood, anxiety, substance use and stress-related disorders, and variability in their reactivity can predict relative risk for illness (Kareken et al., 2004; Jovanovic and Ressler, 2010; Casey et al., 2011; Nikolova et al., 2012), as well as efficacy of common treatments (Bryant et al., 2008; Whalen et al., 2008; Stoy et al., 2012). Thus, developing models that can predict threat-related amygdala and reward-related VS reactivity with high fidelity, without the need for direct assays via neuroimaging, has significant potential to advance ongoing efforts to better treat and possibly prevent mental illness at a population level (Singh and Rose, 2009).

In addition to these fMRI paradigms, each participant completed the NEO-PI-R question-

naire to assess personality across five domains: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. This is a self-report data with 240 total questions, the percentage of missing values is approximately 2%. To simplify the presentation, and because of the nature of this questionnaire, all the responses are ordinal and scored using a 5-point Likert scale (Likert, 1932) ranging from 1 (strongly disagree) to 5 (strongly agree).

Each participant also provided saliva for the identification of single nucleotide polymorphisms from their genomic DNA. Single nucleotide polymorphisms or SNPs represent single points of inter-individual variability in DNA, wherein one nucleotide is substituted with another in a subgroup of individuals, resulting in two different genetic variants or alleles occurring in the population. Each individual's genotype is characterized by two alleles per SNP, thus yielding three possible allele combinations. SNP data (variability in the sequence of A, C, G, T nucleotides) are available for each of the 653 subjects, and in each case we consider a subset of 5568 alleles. This subset is constituted by seeking to concentrate the analysis on SNPs associated with three neurotransmitter systems: norepinephrine (NE), dopamine (DA) and serotonin (5-HT). These three systems were chosen because of their extensive involvement in the regulation of threat- and reward-related brain function (Davis and Whalen, 2001; Schultz, 2005) and known prior associations between genetic variants in those systems and the neural and behavioral traits of interest (Hariri et al., 2002; Nikolova et al., 2011). The genotyping platform contained 794 SNPs for the NE system, 800 SNPs for the DA system, and 1190 SNPs for the 5-HT system (two alleles per SNP). SNP data are generally treated as unordered categorical data (Pritchard et al., 2000; Dunson and Xing, 2009; Bhattacharya and Dunson, 2012). However, an ordinal scale (values 1, 2, 3) can be used to code the SNP genotype sequences under an *additive genetic model* (Elston, 2000). This ordinal scale implies that having two minor (i.e., less common in the population) alleles rather than having no minor alleles is twice as likely to affect the outcome in a certain direction. In this work, we opted for using this ordinal scale for SNP data for modeling purposes.

Finally, we also have access to three additional response variables concerning psychiatric disorders: *internalizing*, *externalizing* and *thought* (Krueger, 1999). Specifically, internalizing disorder is associated with anxious and depression symptoms; externalizing is associated with aggressive, delinquent and hyperactive symptoms as well as alcohol/drug use disorders; and thought disorder is a a disturbance in one's ability to generate a logical sequence of ideas, most commonly associated with schizophrenia or some related psychotic disorder. The real measurements associated with those three psychiatric scores (assessed and scored using a structured clinical interview through the electronic Mini-International Neuropsychi-

atric Interview) are useful to evaluate an individual's propensity to develop multiple forms of psychopathology.

1.3 Neuroscience questions being addressed

Noninvasive human neuroimaging, particularly functional magnetic resonance imaging (fMRI), is routinely employed as an experimental methodology for probing the neurobiological correlates of myriad perceptual, cognitive, and affective processes (Smith, 2012). In addition to its prominent use as a research tool in cognitive neuroscience, fMRI has been increasingly leveraged to better understand the etiology and pathophysiology of mental illness (Korgaonkar et al., 2013). More recently, fMRI measures of brain activity have emerged as particularly important in the context of developing biomarkers that predict relative risk for mental illness and illuminate specific pathways for individually tailored treatment strategies (Singh and Rose, 2009). However, the extensive research infrastructure and associated operating costs of fMRI and related neuroimaging techniques limit the potential extension of such predictive neural biomarkers to clinical settings where the majority of mental health care and treatment is provided. It is thus of interest to develop strategies for accurately predicting individual patterns of behaviorally and clinically relevant brain activity in the absence of direct neuroimaging measures. In this paper we ask the novel and practical question: can one predict neural biomarkers measured with fMRI based solely on information derived from readily administered self-report questionnaires and assayed genetic variation? In another direction, can one predict an individuals fundamental propensity for psychopathology based on observed self-report questionnaires, SNP and fMRI data (separately or in combination)?

1.4 Statistical methods

Multi-view learning seeks to characterize a given entity (here mental health of a particular subject) based upon multiple types of data (or "views"). While some of these methods assume different sensor types are responsible for the data, the *alphabet* of these different views are assumed to be the same (e.g., all views are real valued, categorical or ordinal). Some multi-view methods are targeted toward a specific problem class, such as classification or matrix completion (imputation of missing data). For example, Virtanen et al. (2011); Klami et al. (2013) propose a Bayesian treatment for canonical correlation analysis, to learn statistical relationships between two data sets (views); their approach considers a factor model with a group-wise sparsity prior for the factor loadings. In addition, Virtanen et al.

(2012); Damianou et al. (2012) discuss the choice of sparsity-promoting priors in more detail, and present extensions to more than two views. In Zhe et al. (2014) the authors present a Bayesian approach to identify associations between genetic variations (SNPs) and MRI features and, at the same time, to build a classifier to predict Alzheimer's disease. However, the supervised multi-view learning approach proposed there only assumes two heterogeneous data sources. Alternatively, the joint analysis of ordered, categorical and real data was considered in Salazar et al. (2012, 2013), with an application to cognitive neuroscience data, but those works did not infer relationships between different data sources and used a binary matrix factorization for each view.

This paper addresses analysis of heterogeneous multi-view data, with a neuroscience and mental-health motivation. While fMRI is the particular imaging modality considered here, the same approach may be applied to electroencephalography (EEG) and other brain-imaging modalities (Fyshe et al., 2012). In Stingo et al. (2013) the authors integrated SNP data with clustering of brain regions imaged via fMRI; however, in that study a (relatively small) subset of SNP biomarkers were selected *a priori* as covariates based upon the related literature, and the original SNP data were not analyzed (self-report questionnaire data were also not considered). Finally, data concerning psychiatric disorders, such as internalizing, externalizing and thought are also considered in our general framework.

It is anticipated that the manner in which multiple people answer a given question, or how the brains of multiple people respond to a given stimuli, are *not* independent. While each individual is unique, there are typically statistical relationships between people and their brains, which we wish to infer. Further, for a particular person, the answers to multiple questions are typically not independent, and the fMRI responses to different stimuli are also typically not independent. We wish to infer these statistical relationships, *jointly* across people, questions, brain regions, stimuli, and psychiatric risk scores.

The remainder of the paper is organized as follows. Section 2 describes the generative framework for heterogeneous multi-view data, considering a Bayesian factor model with group-wise sparsity on the factor loadings. Section 3 presents an adaptation for multi-view clustering and factor regression. Section 4 details the variational Bayes algorithm employed to perform approximate inference. Section 5 reports on the principal results and their neuroscience interpretation. In Section 6 we provide summary conclusions from this study. Additional technical details and results are provided in the Supplementary Material.

2 Basic Modeling Framework

2.1 Data and notation

We assume access to data from N people, with the data defined by $M = M_1 + M_2$ heterogeneous data sources (or views). The first M_1 views are assumed to be ordinal-valued data matrices $\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(M_1)}$ and the last M_2 views are assumed to be real-valued data matrices $\boldsymbol{Y}^{(M_1+1)}, \ldots, \boldsymbol{Y}^{(M_1+M_2)}$. Each $\boldsymbol{Y}^{(m)}$ is an $N \times P_m$ matrix, with row i corresponding to P_m data points for view m of person i.

In our motivating neuroscience application, the ordinal-valued views correspond to (i) one response matrix from the NEO-PI-R questionnaire assessing five domains of personality, and (ii) SNP data represented in an ordinal manner (Elston, 2000). For the real-valued data matrices, a subset correspond to fMRI, defined by fMRI data strength in a set of voxels of the brain, in response to a given stimulus. The particular fMRI data matrices are: (i) responses for left and right amygdala reactivity to four visual stimuli, and (ii) reactivity in the left and right ventral striatum (VS). The other type of real-valued data corresponds to scores for psychiatric disorders, specifically *internalizing*, *externalizing* and *thought*.

There is wide variability in the number of data points P_m across the different views: 167 to 329 voxels per brain region for the fMRI data, a self-report questionnire composed of 240 questions, genetic systems ranging from 794 to 1190 SNPs, and 3 psychiatric scores. This wide diversity across the views will impact the form of the model, as discussed below.

We assume that $\boldsymbol{Y}^{(m)}$ is generated via a link function f_m and an underlying matrix $\boldsymbol{X}^{(m)} \in \mathbb{R}^{N \times P_m}$; therefore, $\boldsymbol{Y}^{(m)} = f_m(\boldsymbol{X}^{(m)})$. For real-valued data f_m is the identity link function, so $\boldsymbol{Y}^{(m)} = \boldsymbol{X}^{(m)}$. For ordinal data f_m is the probit link function. Assume $x_{ij}^{(m)}$ is element (i, j) of $\boldsymbol{X}^{(m)}$, with $y_{ij}^{(m)}$ defined in the same manner with respect to $\boldsymbol{Y}^{(m)}$. Further assume that each ordinal view has L_m possible answers, i.e., $y_{ij}^{(m)} \in \{1, \ldots, L_m\}$. The ordered probit model is defined by $y_{ij}^{(m)} = l \in \{1, \ldots, L_m\}$ if $g_{l-1}^m < x_{ij}^{(m)} < g_l^m$, with cut-points $g_0^m = -\infty$, $g_{L_m}^m = \infty$, and $g_1^m < \ldots < g_{L_m-1}^m$. The parameters of this model are the real latent matrix $\boldsymbol{X}^{(m)}$ and the cut-points $\{g_1^m, \ldots, g_{L_m-1}^m\}$.

2.2 Joint modeling for heterogeneous multi-view data

Let vector $\boldsymbol{x}_i^{(m)}$ represent row i of $\boldsymbol{X}^{(m)}$, which we model as

$$\boldsymbol{x}_i^{(m)} \sim \mathcal{N}(\boldsymbol{v}_i^{(m)} \boldsymbol{W}^{(m)}, \gamma_m^{-1} \boldsymbol{I}), \quad \boldsymbol{v}_i^{(m)} \sim \mathcal{N}(\boldsymbol{v}_i, \tau_m^{-1} \boldsymbol{I}), \quad \boldsymbol{v}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{1}$$

with $\boldsymbol{v}_i^{(m)} \in \mathbb{R}^{1\times K}$, $\boldsymbol{v}_i \in \mathbb{R}^{1\times K}$, $\boldsymbol{W}^{(m)} \in \mathbb{R}^{K\times P_m}$ and $K << \sum_{m=1}^{M} P_m$ denotes the number of latent factors. For the precision parameter γ_m we assign an uninformative Gamma prior, $\gamma_m \sim \text{Ga}(a_\gamma, b_\gamma)$, where the hyperpriors a_γ, b_γ are set to small values. In (1) τ_m is a view-specific precision parameter, such that $\tau_m \sim \text{Ga}(a_\tau, b_\tau)$ with hyperpriors a_τ, b_τ set to small values. This model allows view-specific factor scores $\boldsymbol{v}_i^{(m)}$, each of these varying about a mean factor score $\boldsymbol{v}_i$ for subject i. Note that when $\tau_m \to \infty$ for all m, $\boldsymbol{v}_i^{(m)} \to \boldsymbol{v}_i$, and (1) reduces to the simpler model for which all the factor scores $\boldsymbol{v}_i^{(m)}$ are the same for all m; such models were considered in (e.g. Jia et al., 2010; Chen et al., 2010; Virtanen et al., 2012; Klami et al., 2013; Zhe et al., 2014). After introducing our prior on the factor loadings $\boldsymbol{W}^{(m)}$, we revisit the factor-score model, and make connections to previous work. We also provide further explanation for why this model (with use of τ_m) is needed for the data considered.

2.3 Block-wise sparsity

We use group-wise automatic relevant determination (ARD) (Neil, 1996; Tipping, 2001) as the sparsity-inducing prior on $\{\boldsymbol{W}^{(m)}\}_{m=1}^{M}$, which also helps infer the number of latent factors by shrinking the unnecessary rows in each $\boldsymbol{W}^{(m)}$ to zero (hence K is an upper bound on the number of factors, and the data are used to infer the number – typically less than K – of factors actually needed). Let $\boldsymbol{w}_k^{(m)}$ represent row k of $\boldsymbol{W}^{(m)}$, modeled as

$$\boldsymbol{w}_k^{(m)} \sim \mathcal{N}(\boldsymbol{0}, \alpha_{mk}^{-1} \boldsymbol{I}), \quad \alpha_{mk} \sim \text{Ga}(a_\alpha, b_\alpha). \tag{2}$$

The hyperpriors a_α and b_α are set to small values to favor noninformative priors with wide support. Note that, by choosing $a_\alpha, b_\alpha \to 0$ we obtain the Jeffrey's prior $p(\alpha_{mk}) \propto 1/\alpha_{mk}$, favoring strong sparsity.

The model in (1) may be rewritten as

$$\boldsymbol{x}_i^{(m)} \sim \mathcal{N}(\boldsymbol{v}_i \tilde{\boldsymbol{W}}^{(m)}, \gamma_m^{-1} \boldsymbol{I}), \quad \tilde{\boldsymbol{w}}_k^{(m)} \sim \mathcal{N}(\boldsymbol{0}, \tilde{\alpha}_{mk}^{-1} \boldsymbol{I}), \quad \boldsymbol{v}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{3}$$

where $\tilde{\alpha}_{mk}^{-1} = (1+\tau_m^{-1})\alpha_{mk}^{-1}$. Making connections to the model in (2), factor loadings $\tilde{\boldsymbol{w}}_k^{(m)}$ and

$\boldsymbol{w}_k^{(m)}$ are related as $\tilde{\boldsymbol{w}}_k^{(m)} = \sqrt{1+\tau_m^{-1}}\boldsymbol{w}_k^{(m)}$; the introduction of τ_m is equivalent to adding the flexibility of view-specific scaling of the factor loadings. With comparison to (1), the model in (3) employs factor scores $\boldsymbol{v}_i$ that are independent of the view m, with the view-specific scaling now absorbed in the factor loadings $\{\tilde{\boldsymbol{W}}^{(m)}\}_{m=1}^M$. If τ_m^{-1} tends towards zero, then $\tilde{\boldsymbol{w}}_k^{(m)} \to \boldsymbol{w}_k^{(m)}$ and $\tilde{\alpha}_{mk}^{-1} \to \alpha_{mk}^{-1}$. We can identify the factor activeness in each view from the precision $\tilde{\alpha}_{mk}$. Factor k is inactive in view m if $\tilde{\alpha}_{mk}$ becomes large (i.e., $\tilde{\alpha}_{mk}^{-1} \to 0$). The posterior distributions over the $\tilde{\alpha}_{mk}$ help to select the number of components (or factors) needed for each view. The kth factor becomes inactive when $\tilde{\alpha}_{mk}$ has large values for all views $m \in \{1,\ldots,M\}$. The kth factor may also only contribute to a *subset* of views, those views m for which $\tilde{\alpha}_{mk}$ is large. We emphasize that the elements of $\tilde{\boldsymbol{w}}_k^{(m)}$ for inactive views are not exactly zero, but instead they are pushed to very small values (near-sparsity).

In the course of analyzing the motivating multi-view neuroscience data, we first considered a model with $\tau_m \to \infty$, analogous to Jia et al. (2010); Chen et al. (2010); Virtanen et al. (2012); Klami et al. (2013); Zhe et al. (2014); the results (e.g., prediction accuracy) were significantly inferior to those produced by the proposed model. Letting $\tau_m \to \infty$ corresponds to sharing the factor scores $\boldsymbol{v}_i$ across views, and imposing a fixed form of shrinkage (2) on the factor loadings. The model in (3) shares the factor scores $\boldsymbol{v}_i$ but $\tilde{\alpha}_{mk}^{-1} = (1+\tau_m^{-1})\alpha_{mk}^{-1}$ allows a more-flexible, view-dependent scaling of the precision on the factor loadings, found necessary for the highly imbalanced and heterogeneous data considered. This representation for $\tilde{\alpha}_{mk}^{-1}$ is reminiscent of the local-global shrinkage priors developed in Polson and Scott (2010). A fixed prior is placed on all α_{mk}, for each of the "local" components k, and then a scaling via τ_m is effected "globally" across all components k associated with view m; the global scaling is view dependent. This local-global setup plays an important role for the heterogeneous and imbalanced multi-view data considered here.

We underscore that precision $\tilde{\alpha}_{mk}$ in (3) applies to all P_m data components in view m, as modeled by factor k. So the shrinkage in the factor loadings is employed at the *block level*, where block m corresponds to the P_m data points in that view (e.g., the answers from the NEO personality test, SNP data, or fMRI voxel values in a brain region in response to visual stimuli). Consequently, the prior imposes that a given factor is likely to be important (small $\tilde{\alpha}_{mk}$) or unimportant (large $\tilde{\alpha}_{mk}$) for all or most data in a given view, and some factors may be specialized to a subset of views.

Correlations between views m and m' can be computed via the view-specific factor loading matrices $\{\tilde{\boldsymbol{W}}^{(m)}\}_{m=1}^M$. Let $\boldsymbol{C} = \{C_{m\,m'}\} \in \mathbb{R}^{M\times M}$ be the view-correlation matrix. Then,

$C_{m\,m'} = (\bar{\boldsymbol{w}}^{(m)})^\top \bar{\boldsymbol{w}}^{(m')}/(s^{(m)}s^{(m')})$ where $\bar{\boldsymbol{w}}^{(m)} = (\bar{w}_1^{(m)}; \ldots; \bar{w}_K^{(m)})$, $\bar{w}_k^{(m)} = \sum_{j=1}^{P_m} (\tilde{w}_{kj}^{(m)})^2$ (for $k = 1, \ldots, K$) and $s^{(m)} = \sqrt{(\bar{\boldsymbol{w}}^{(m)})^\top \bar{\boldsymbol{w}}^{(m)}}$.

2.4 Identifiability via rotation

The multi-view factor model specified in (3) is in general unidentifiable due to the fact that $\boldsymbol{V}\tilde{\boldsymbol{W}}^{(m)} = \boldsymbol{V}\boldsymbol{Q}\boldsymbol{Q}^{-1}\tilde{\boldsymbol{W}}^{(m)}$, for arbitrary rotation $\boldsymbol{Q}$; the set of factor-score vectors $\{\boldsymbol{v}_i\}$ define the rows of $\boldsymbol{V}$. To solve this problem, we follow the idea of parameter-expanded variational Bayes (Qi and Jaakkola, 2007; Luttinen and Ilin, 2010; Virtanen et al., 2012) and derive a more efficient algorithm for optimizing the variational approximation. That is, we explicitly optimize w.r.t. $\boldsymbol{Q}$ to maintain identifiability in the model, and achieve faster convergence during inference.

3 Model Extensions

3.1 Multi-view clustering

The mean factor scores associated with the N subjects, $\{\boldsymbol{v}_i\}$, have thus far been modeled independently. It is anticipated that people may cluster into types, and this may be leveraged when learning the factor scores. We assume that the $\{\boldsymbol{v}_i\}$ are generated from a J component Gaussian mixture model (GMM) such that, for $j = 1, \ldots, J$

$$\boldsymbol{v}_i|z_i, \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{I}), \quad z_i|\boldsymbol{\pi} \sim \text{Discrete}(\boldsymbol{\pi}), \quad \boldsymbol{\mu}_j \sim G_0, \quad \boldsymbol{\pi}|\alpha \sim \text{Dirichlet}(\alpha/J, \ldots, \alpha/J), \quad (4)$$

where z_i denotes the cluster assignment for the subject i, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$ is the vector of mixing proportions ($\sum_{j=1}^J \pi_j = 1$), $\boldsymbol{\mu}_j$ is the mean vector for component j, and G_0 is the prior distribution on $\boldsymbol{\mu}_j$ (e.g., G_0 may be $\mathcal{N}(\boldsymbol{0}, \eta^{-1}\boldsymbol{I})$). This construction can be generalized by replacing the finite GMM by a Dirichlet Process mixture model (DPMM) (Escobar and West, 1995), which corresponds to $J \to \infty$ in (4) (with the DPMM one may infer the number of needed mixture components, rather than setting a truncation J). In the experiments presented below, the simple finite GMM, with relatively large J, was found to work well (with the finite GMM, one may adjust J to be large enough such that only a subset of mixture components have appreciable posterior probability of being used).

The above multi-view clustering framework is appealing because it handles views having diverse types of representations, allows data to be missing in one or more views, and

performs *simultaneous* factor modeling and clustering. In contrast, existing multi-view clustering methods (such as Bickel and Scheffer, 2004; Chaudhuri et al., 2009; Kumar et al., 2011; Wang et al., 2013) have one or more of the following limitations: (*i*) all views are assumed to have the same representation (real-valued feature or kernel matrices), (*ii*) all views contribute equally/similarly to the global factor representation of the data, and (*iii*) the clustering is performed in a *post-hoc* fashion (Chaudhuri et al., 2009; Kumar et al., 2011) after the global factor representation has been learned. The clustering adaptation of the proposed framework does not have any of these limitations.

3.2 Predictive factor regression model

Assume that the goal is to predict $\boldsymbol{x}_i^{(m)} \in \mathbb{R}^{P_m}$ for real-valued view m, based upon all or a subset of $\boldsymbol{y}^{(m')}$ for $m' \neq m$; we will be interested in doing this when predicting the three psychiatric scores, as well as when predicting fMRI readings. From (3) we have

$$\boldsymbol{x}_i^{(m)} \sim \mathcal{N}(\boldsymbol{v}_i \tilde{\boldsymbol{W}}^{(m)} + \boldsymbol{c}_i \boldsymbol{B}^{(m)}, \gamma_m^{-1} \boldsymbol{I}). \tag{5}$$

where $\boldsymbol{B}^{(m)} \in \mathbb{R}^{R \times P_m}$ is a view-specific matrix added to the model when one has access to covariates $\boldsymbol{c}_i \in \mathbb{R}^R$ for individual i. For each row of $\boldsymbol{B}^{(m)}$, we use an ARD prior to impose group-wise sparsity, as described in Section 2.3.

We learn distributions for $\{\tilde{\boldsymbol{W}}^{(m)}\}$, $\{\boldsymbol{B}^{(m)}\}$ and $\{\gamma_m\}$ based on training data. When performing a prediction of a particular $\boldsymbol{x}_i^{(m)}$ based on $\boldsymbol{D}$, where $\boldsymbol{D}$ is a selected set of data $\boldsymbol{y}_i^{(m')}$ for $m' \neq m$, we estimate a posterior distribution for $p(\boldsymbol{v}_i|\boldsymbol{D})$. Using $p(\boldsymbol{v}_i|\boldsymbol{D})$ and the distributions for $\{\tilde{\boldsymbol{W}}^{(m)}\}$, $\{\boldsymbol{B}^{(m)}\}$ and $\{\gamma_m\}$, one may predict (e.g., the mean) missing $\boldsymbol{x}_i^{(m)}$ via (5).

The complete model is summarized in Figure 2. It may be viewed as an extension of the sparse latent factor regression model in West (2003) to a multi-view setting; Carvalho et al. (2008) also propose a similar framework but without considering a group-wise sparsity on the factor loadings, because multi-view data was not considered.

4 Posterior Inference

We seek to infer the posterior distribution of all latent parameters, given the data. Several computational methods may be considered, such as the variational Bayesian algorithm (Beal,

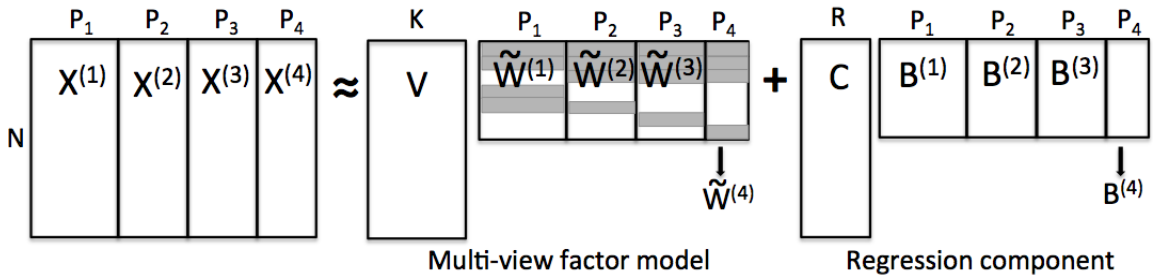


Figure 2: Graphical representation of the multi-view factor model with regression component with $M = 4$ views. Here, the $N \times K$ matrix $\boldsymbol{V}$ represents the *global* latent representation of the data. The factor loading matrix shows a group-wise sparsity pattern, where some of the factors are active in all the views whereas others are only active in some subset of views. Gray areas represent activeness of factor k in view m. The regression component is given by the product of the matrix of known covariates $\boldsymbol{C}$ and the view-specific matrix of regression coefficients.

2003) and Markov chain Monte Carlo (MCMC) with Gibbs sampling (Gelfand and Smith, 1990). For multi-view learning Virtanen et al. (2012), Klami et al. (2013) and Zhe et al. (2014) derive a variational mean-field algorithm (with efficient ARD prior updates). In Klami and Kashi (2007) a Gibbs sampling algorithm is derived (considering only two views but easily extended for more than two) to achieve an approximation to the posterior distribution. However, in general, Gibbs sampling is inefficient for large-dimensionality problems (i.e., for large P_m, M and N). Given the large size of the data considered here, we employ variational Bayesian (VB) inference, with key aspects of the VB implementation described in this section. In addition, the Supplementary Material also contains further details about an MCMC implementation, to which we compared our VB results.

Without loss of generality, throughout this section we only consider inference for the model in (3) (i.e., without covariates). Inference for the view-specific regression coefficient parameters (for covariates) is straightforward to implement and follows the same update equations of the view-specific factor loading matrix, with some small modifications. We infer the variational distribution for the latent variables, collectively referred to as $\boldsymbol{\Theta}$, and consisting of $\{\tilde{\boldsymbol{W}}^{(m)}, \boldsymbol{\alpha}_m, \gamma_m, \tau_m\}_{m=1}^M, \boldsymbol{V}\}$, along with $\{\{\boldsymbol{\mu}_j\}_{j=1}^J, \boldsymbol{z}\}$ for clustering. For the cutpoints $\boldsymbol{G} = \{\boldsymbol{g}^m\}_{m=1}^{M_1}$ and the rotation matrix $\boldsymbol{Q}$, we seek a point estimate. Sometimes, for brevity, we will use $\{\tilde{\boldsymbol{W}}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\tau}\}$ for $\{\tilde{\boldsymbol{W}}^{(m)}, \boldsymbol{\alpha}_m, \gamma_m, \tau_m\}_{m=1}^M$ and $\{\boldsymbol{\mu}, \boldsymbol{z}\}$ for $\{\{\boldsymbol{\mu}_j\}_{j=1}^J, \boldsymbol{z}\}$, respectively. The data from all views are collectively referred to as $\boldsymbol{\mathcal{Y}}$. Here we only provide brief descriptions of the key aspects of our inference algorithm. The Supplementary Material contains further details, including the update equations for the clustering adaptation.

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

$$q(\boldsymbol{\Theta}) = \prod_{i=1}^{N} q(\boldsymbol{v}_i) \prod_{m=1}^{M} \left(\prod_{k=1}^{K} q(\tilde{\boldsymbol{w}}_k^{(m)}) \prod_{k=1}^{K} q(\alpha_{mk}) q(\gamma_m)\ q(\tau_m) \right). \tag{6}$$

Thus, we minimize the KL-divergence $KL(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}|\boldsymbol{\mathcal{Y}}, \boldsymbol{G}, \boldsymbol{Q}))$, which is equivalent to maximizing the evidence lower bound (ELBO) given by $\mathcal{L}(q(\boldsymbol{\Theta}), \boldsymbol{G}, \boldsymbol{Q}) = \mathbb{E}_{q(\boldsymbol{\Theta})}[\log p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\Theta}|\boldsymbol{G}, \boldsymbol{Q}) - \log(q(\boldsymbol{\Theta}))]$. With further approximation for the ordinal-valued views using a Taylor-series expansion, we efficiently update variational parameters for $q(\boldsymbol{\Theta})$. Note that in (6), terms associated with the latent variables $\boldsymbol{x}_i^{(m)}$ (for ordinal views) do not appear; that is because for ordinal-based views the $\boldsymbol{x}_i^{(m)}$'s are integrated out in the variational lower bound (details of derivation are provided in the 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 $\boldsymbol{V}$, $\tilde{\boldsymbol{W}}$, the cutpoints $\boldsymbol{G}$ and the rotation matrix $\boldsymbol{Q}$.

Updating $\boldsymbol{v}_i$: $q(\boldsymbol{v}_i) = \sum_{c=1}^{J} q(z_i = c)\mathcal{N}(\boldsymbol{\mu}_{v,c}, \boldsymbol{\Sigma}_{v,c})$. For the multi-view clustering case with GMM prior on $\boldsymbol{v}_i$, we have that $\boldsymbol{\Sigma}_{v,c} = (\boldsymbol{I} + \sum_{m=1}^{M} \langle \gamma_m \rangle \langle \tilde{\boldsymbol{W}}^{(m)} \tilde{\boldsymbol{W}}^{(m)\top} \rangle)^{-1}$ and $\boldsymbol{\mu}_{v,c} = (\boldsymbol{\mu}_c + \sum_{m=1}^{M_1} \langle \gamma_m \rangle \sum_{j=1}^{P_m} \langle \tilde{\boldsymbol{W}}_{:j}^{(m)} \top \rangle (\boldsymbol{g}^m_{y_{i,j}^{(m)}} + \boldsymbol{g}^m_{y_{i,j}^{(m)}-1})/2 + \sum_{m=M_1+1}^{M} \langle \gamma_m \rangle \boldsymbol{y}_i^{(m)} \langle \tilde{\boldsymbol{W}}^{(m)\top} \rangle$. Such that $\langle V_i \rangle = \sum_{c=1}^{J} q(z_i = c)\boldsymbol{\mu}_{v,c}$ and $\langle \boldsymbol{v}_i^\top \boldsymbol{v}_i \rangle = \sum_{c=1}^{J} q(z_i = c)(\boldsymbol{\mu}_{v,c}^\top \boldsymbol{\mu}_{v,c} + \boldsymbol{\Sigma}_{v,c})$.

Updating $\tilde{\boldsymbol{W}}^{(m)}$: The variational posterior for the j-th column of $\tilde{\boldsymbol{W}}^{(m)}$ $(j = 1, \ldots, P_m)$, is given by $q(\tilde{\boldsymbol{W}}_{:j}^{(m)}) = \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$, where $\boldsymbol{\Sigma}_w = (\langle \text{diag}(\tilde{\alpha}_{m1}, \ldots, \tilde{\alpha}_{mK}) \rangle + \langle \gamma_t \rangle \sum_{i=1}^{N} \langle \boldsymbol{v}_i^\top \boldsymbol{v}_i \rangle)^{-1}$ and $\boldsymbol{\mu}_w$ depends on the type of view. For ordinal-valued feature-based view, $\boldsymbol{\mu}_w = \boldsymbol{\Sigma}_w \langle \gamma_m \rangle \sum_{i=1}^{N} \langle \boldsymbol{v}_i^\top \rangle (\boldsymbol{g}^m_{y_{i,j}^{(m)}} + \boldsymbol{g}^m_{y_{i,j}^{(m)}-1})/2$. For real-valued feature-based view, $\boldsymbol{\mu}_w = \boldsymbol{\Sigma}_w \langle \gamma_m \rangle \sum_{i=1}^{N} \langle \boldsymbol{v}_i^\top \rangle y_{ij}^{(m)}$.

Inferring the cutpoints: We infer the cut-points $\boldsymbol{g}^m$ 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 $\tilde{\mathcal{L}}_l^m = N_l^m[\log(\boldsymbol{g}_l^m - \boldsymbol{g}_{l-1}^m) - \langle \gamma_m \rangle (\boldsymbol{g}_l^{m2} + \boldsymbol{g}_{l-1}^{m\,2} + \boldsymbol{g}_l^m \boldsymbol{g}_{l-1}^m)/6] + \langle \gamma_m \rangle (\boldsymbol{g}_l^m + \boldsymbol{g}_{l-1}^m) \sum_{i,j:y_{i,j}^m = l} \langle \boldsymbol{v}_i \rangle \langle \tilde{\boldsymbol{W}}_{:j}^{(m)} \rangle / 2$. Here N_l^m is the number of observations with value l in view m. Also, we set $\boldsymbol{g}_0^m = -G$ and $\boldsymbol{g}_{L_m}^m = +G$ with G fixed and large. The gradients of $\tilde{\mathcal{L}}_l^m$ 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 $p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\Theta})$ 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 step encourages structured sparsity (imposed by prior), helps escaping from local optima, and achieves a faster convergence (Virtanen et al., 2012). The rotation matrix also helps in the identifiability of the inferred loading matrices.

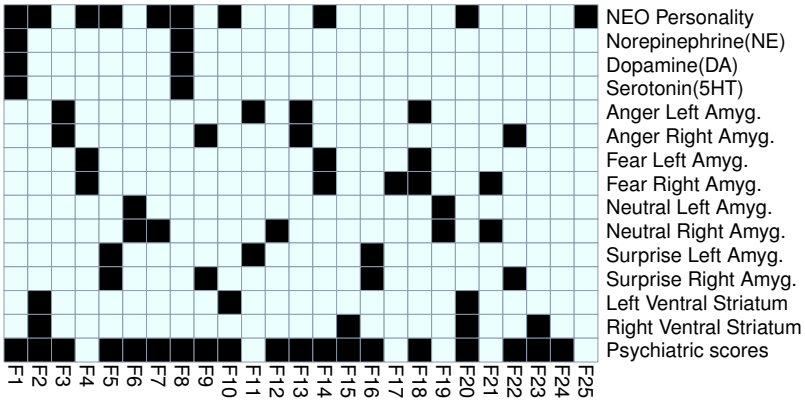
5 Experimental Results

We apply the multi-view factor modeling framework on clinically-relevant neuroscience data, to: (i) interpret common/view-specific latent factors as well as view-correlations, (ii) predict missing data for some views (e.g., fMRI responses and psychiatric scores) leveraging information from multiple views, (iii) cluster people. The heterogeneous multi-view data were collected from 653 college students and consist of $M = 15$ views, the first $M_1 = 4$ views are ordinal-valued matrices and the last $M_2 = 11$ views are real-valued matrices. Specifically, we consider 1 ordinal-valued response matrix from the NEO-PI-R questionnaire; 3 ordinal-valued views from SNP data (from the NE, DA, 5-HT systems); 8 real-valued views from left/right amygdala reactivity to anger/fear/neutral/surprise stimuli; and 2 real-valued views from left/right VS. We also considered gender, which has demonstrated effects on both our NEO-PI-R and fMRI measures, as a covariate in the regression component to fit the model.

We perform analysis considering an upper bound of $K = 50$ latent factors, and prior hyperparameters $a_\alpha = b_\alpha = a_\tau = b_\tau = 0.01$. We noticed that K is large enough, since the number of *active* factors in the model is around 25. All experiments were performed using MATLAB code. The VB based inference method converged (in terms of the variational lower bound) in about 25 iterations. Comparison with MCMC-based results are provided in Supplementary Material.

5.1 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. For instance, we can identify the factor activeness in each view from the precision $\tilde{\alpha}_{mk}$. That is, factor k is inactive in view m if $\tilde{\alpha}_{mk}$ is relatively large (i.e., $\tilde{\alpha}_{mk}^{-1} \to 0$). In order to identify active factor across all views, let $\boldsymbol{B} = \{b_{mk}\}$ be an $M \times K$ binary matrix, where the non-zero elements indicate dependency between active views at each factor. We set $b_{mk} = 1$ if $\tilde{\alpha}_{mk}^{-1} > \epsilon$, for



	Conscientiousness	Extraversion	Openness	Agreeableness	Neuroticism
F1	20.7%	20.7%	18.2%	18.2%	22.2%
F2	***45.4%***	17.6%	17.6%	8.5%	11.8%
F4	7.2%	***63.4%***	6.1%	5.2%	18.1%
F5	10.5%	***7.6%***	***43.5%***	32.6%	5.8%
F7	7.2%	9.4%	***62.7%***	10.5%	10.2%
F8	16.3%	14.7%	24.8%	24.8%	19.4%
F10	15.7%	14.3%	3.6%	***55.0%***	11.4%
F14	6.1%	13.3%	9.1%	7.3%	***64.2%***
F20	15.8%	26.3%	18.4%	18.4%	21.1%
F25	22.1%	21.5%	22.7%	17.7%	16.0%

Figure 3: **Left**: Matrix with elements $\tilde{\alpha}_{mk}^{-1} > 0.01$ ($m = 1, \ldots, 15$; $k = 1, \ldots, 25$) indicating active views for each factor. Row labels indicate the type of view, column labels index factors. **Right**: For the NEO-PI-R questionnaire and for each one of the 10 active factors (first row in the left panel), percentage of questions assessing each of the five personality traits: conscientiousness, extraversion, openness, agreeableness, and neuroticism.

some small ϵ (e.g., 0.01). Figure 3 (left panel) shows the view-factor association matrix $\boldsymbol{B}$ for this data. We only show 25 factors which have at least one active view. Note that there are some factors that have only one/few active views. Further insights can be obtained by interpreting the factor loadings $\tilde{\boldsymbol{W}}^{(m)}$ (for which the rows correspond to factors and columns to questions/SNPs/fMRI voxels). For instance, the NEO-PI-R questionnaire (with 240 questions) is of particular interest in psychology to measure the five broad domains of personality: openness, conscientiousness, extraversion, agreeableness, and neuroticism. Thus, for the NEO-PI-R view and for the 10 active factors therein, we could identify questions that have *substantial loadings* on a given factor by inspecting their factor loading scales; with "substantial" quantified as loadings greater than a threshold value (e.g., for factor k and question j, $|\tilde{w}_{kj}| > 0.1$)[1]. Figure 3 (right panel) shows the percentage of questions associated with every domain of personality (conscientiousness, extraversion, openness, agreeableness, neuroticism). It is insightful to observe that some factors (1st, 20th and 25th) include, in an equitable manner, questions related with the five domains, whereas for other factors, questions related with one or two domains of personality are dominant. These findings are useful for interpreting and naming a factor and to establish some connections between personality domains and the other views.

Figure 4 (left panel) shows the view-correlation matrix inferred from $\tilde{\boldsymbol{W}}^{(m)}$, computed as described in Section 2 (a graphical representation is shown in the right panel, considering correlations greater than or equal to 0.3). As the figure shows, our model discovers views

[1]The 0.1 threshold value represents the 10th percentile of the absolute loadings.

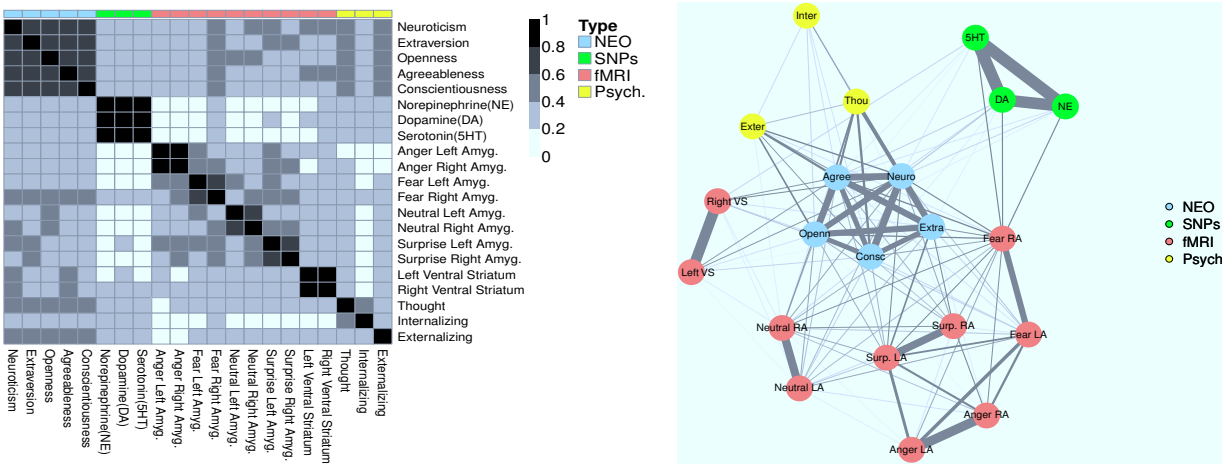


Figure 4: **Left**: Inferred view-correlation matrix. **Right**: Graph representation of the view-correlation matrix (correlations greater than or equal to 0.3 are highlighted) where every node represent a view. Nodes are colored according to the type of data. The view associated with the NEO-PI-R questionnaire was split into five sub-views (or domains) of personality: neuroticism, extraversion, openness, agreeableness and conscientiousness. Also, the view for psychiatric scores was split into thought, internalizing and externalizing disorders.

that have high pairwise correlations. For instance, the correlation matrix reveals novel and unexpected associations between the five domains of personality assessed by the NEO-PI-R questionnaire, genetic data (SNPs in NE, DA and 5HT) and brain activity.

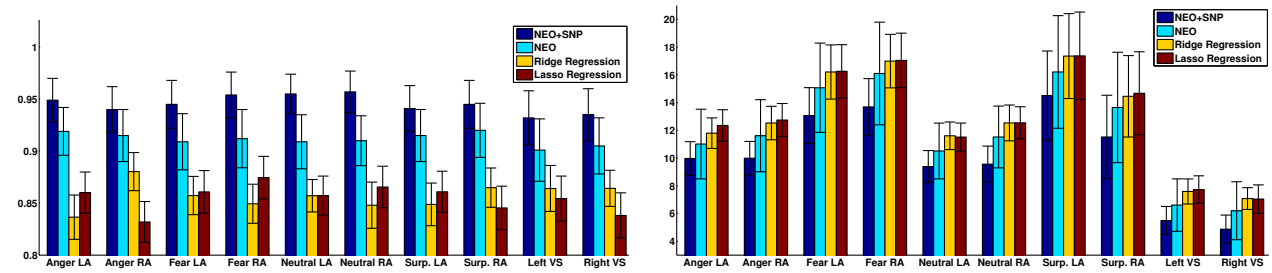
Specifically, the view correlation matrix indicates that personality assessed using the NEO-PI-R is more strongly associated with thought and externalizing disorders than internalizing disorders. The data generated from our framework thus suggests that common measures of personality may be more useful in identifying risk for externalizing and thought disorders. Additional research, including replication in independent and varied samples, is needed to further evaluate the utility of this observation. Another interesting pattern revealed from the view-correlation matrix is the somewhat unique correlation of internalizing disorders with right amygala reactivity to fear. This is particularly noteworthy given the core symptoms of increased sensitivity to threat and trait anxiety common across internalizing disorders, and the importance of the right amygdala in generating phasic responses to threat-related stimuli, especially those of an ambiguous nature as exemplified by the fearful facial expressions used on our fMRI task.

5.2 Predicting fMRI responses and psychiatric scores

The machines used to measure fMRI data are expensive, and require expertise to operate. One of the challenges we wish to examine is whether one can predict fMRI responses to relevant stimuli, based only on how a subject answers questionnaires (which are of course much less costly to administer). In some cases the answers to questionnaires may be combined with SNP data for prediction of fMRI data. SNP data may be constituted from multiple biological tissues including, blood, somatic cells, and saliva, the latter of which was the source of DNA in our sample. While the technology for extracting the SNP data from such a sample is sophisticated, there are companies available for this purpose. Consequently, SNP data are generally more accessible than fMRI, although the analysis cost may be too high currently for routine use.

In this study, we omit fMRI data from 30% of the subjects, i.e., we consider $N_L = 457$ subjects for model learning and $N_T = 196$ subjects for testing. We perform Monte Carlo simulation based on 50 runs (randomly assigning individuals to the learning/testing sets). For the testing group, we only assume access to the ordinal-based views, i.e., the NEO-PI-R questionnaire and SNP data. For model comparison, we consider two baseline models: (1) a ridge regression where the covariates are the ordinal responses; and (2) a Lasso regression model (Tibshirani, 1996) with the same covariates. As comparison criteria, we use the root mean square error (RMSE) and the coefficient of determination (R^2). Specifically, RMSE $= \sqrt{\sum_{i=1}^{N_T}\sum_{j=1}^{P_m}(y_{ij}^{(m)} - \hat{y}_{ij}^{(m)})^2/(N_T P_m)}$ and $R^2 = 1 - \sum_{i=1}^{N_T}\sum_{j=1}^{P_m}(y_{ij}^{(m)} - \hat{y}_{ij}^{(m)})^2 / \sum_{i=1}^{N_T}\sum_{j=1}^{P_m}(y_{ij}^{(m)} - \bar{y}^{(m)})^2$, such that $\hat{y}_{ij}^{(m)} = \mathbb{E}(y_{ij}^{(m)})$ and $\bar{y}^{(m)} = \sum_{i=1}^{N_T}\sum_{j=1}^{P_m} y_{ij}^{(m)}/(N_T P_m)$. The coefficient of determination - R^2, which ranges between 0 and 1, is used to show how accurately a given model can predict unobserved outcomes. We evaluate the prediction performance of the fMRI data based on NEO-PI-R responses alone, SNPs alone, and the combination of NEO-PI-R responses and SNPs. For each of the 10 views associated with the fMRI data (left/right amygdala for four expressions and left/right VS), Figure 5 shows the average of the R^2 and RMSE calculated over 50 runs. In the figure, prediction results based upon NEO-PI-R responses alone and SNPs alone are not shown because they are worse than with the baseline models and the multi-view model with different view combinations (i.e., R^2 lower than 0.55 and RMSE greater than 9).

Here, the SNP data considered were from the three monoamine networks discussed in Section 1.2, known to be important in modulating brain function, including amygdala and VS reactivity. However, we also considered randomly selected sets of alleles, and the predictive performance was virtually unchanged. From these results, we note that the inclusion of



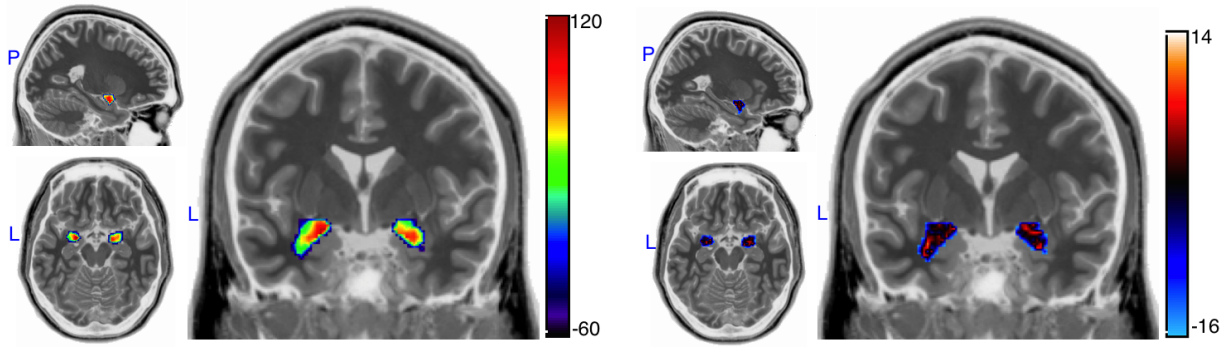
(a) R^2 (b) RMSE

Figure 5: Average of the coefficient of determination - R^2 (a) and root mean square error - RMSE (b) for fMRI and VS predictions calculated over 50 runs. Vertical lines represent the standard deviations. The results marked NEO and NEO+SNP are based on the proposed model (different combinations of data), and the Ridge and Lasso Regression results are based on NEO and SNP data concatenated.

SNP data yields better predictions of fMRI data. Also, the multi-view modeling framework, in general, outperforms the other baselines, showing the benefits of leveraging information between multiple views.

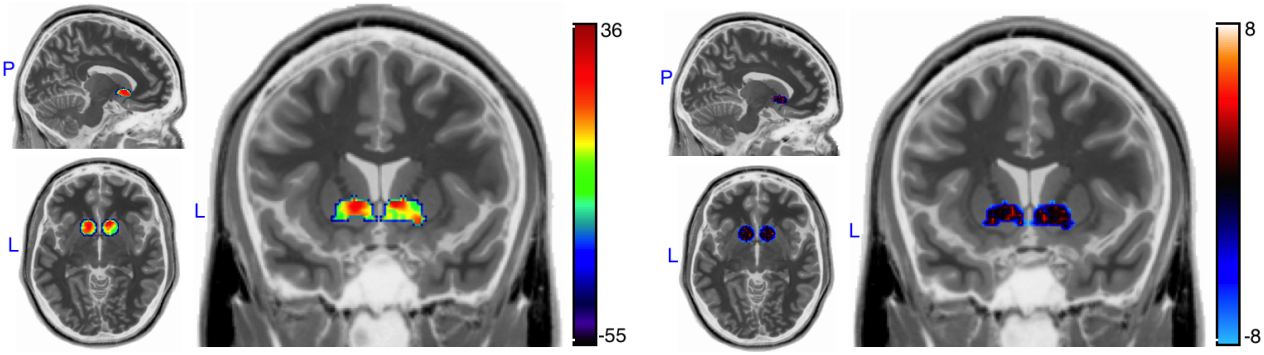
To give a visual sense of the accuracy with which the fMRI data may be predicted, in Figure 6, we show the posterior mean predicted values and prediction errors for left and right amygdala reactivity to fear full expressions, corresponding to one subject in the testing set. The predictions are based on NEO-PI-R responses and SNP data. Figure 7 shows similar results for left and right VS predictions. The figures give us a sense of the 3-dimensional pattern of the predictions. Specifically, for left and right amygdala we predict amygdala reactivity to fear in 167 and 194 voxels, respectively; whereas for left and right VS we predict reactivity in 301 and 329 voxels, respectively.

Our next experiment considers predicting the three psychiatric scores (internalizing, externalizing, and thought). To do that, we include psychiatric score data as another view in the model. As before, when performing learning we omit data from 30% of the subjects (testing set). For that group, we only assume access to the NEO-PI-R responses, SNP and fMRI data. We compare the results with two baselines models (as before) and assess the prediction performance of the psychiatric scores based on NEO-PI-R responses alone, fMRI data alone, SNPs alone, and the combination of the three above. Figure 8 shows the average of the R^2 and RMSE calculated over 50 runs for the two baselines models, and for predictions based on some data-input scenarios, for which we obtained the better results. From these results, we note that the best prediction performance is obtained when the NEO-PI-R, fMRI and SNP views are used in the model. For each of the predicted psychiatric scores, the



(a) Predicted values (b) Prediction error

Figure 6: Sagittal, transverse and coronal views of predicted values (a), and prediction errors (b) for left/right amygdala reactivity to fear (values scaled by 100). Brain images correspond to one subject.



(a) Predicted values (b) Prediction error

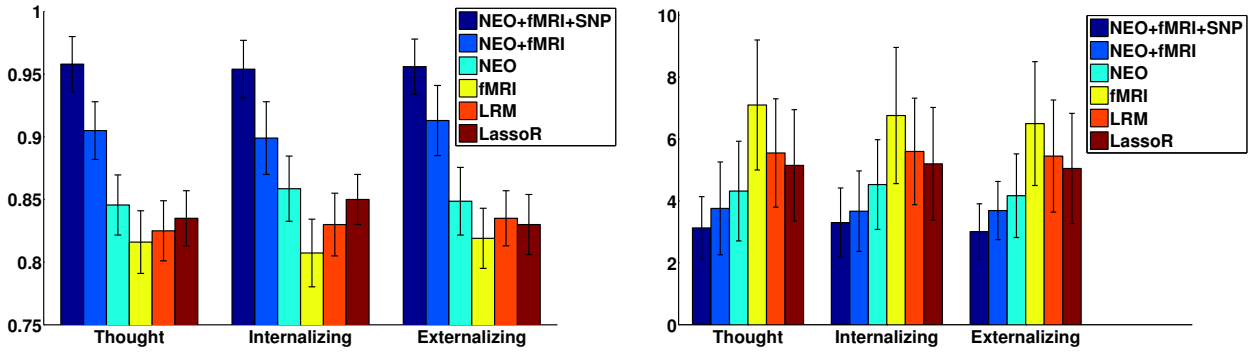
Figure 7: Sagittal, transverse and coronal views of predicted values (a), and prediction errors (b) for left/right ventral striatum (values scaled by 100). Brain images correspond to one subject.

smallest RMSE and the largest R^2 is associated with the multi-view modeling framework with NEO-PI-R, fMRI and SNP views.

Additional results for ordinal matrix completion as well as results considering more questionnaires to fit the model are provided in the Supplementary Material.

5.3 Results for multi-view clustering

We used the clustering framework in Section 3.1 to cluster subjects into groups. Specifically, we use a truncated mixture model with parameters $\alpha = 1$ and $\eta = 1$ and consider all the available views to fit the model. The approximate VB algorithm uses a truncated distribution with a fixed maximum number of components. We considered an upper bound of $J = 25$, and five mixture components fit the data well (5 of the 25 mixture components had nonnegligible posterior probability of being utilized).



(a) R^2 (b) RMSE

Figure 8: Average of the coefficient of determination - R^2 (a) and root mean square error - RMSE (b) for psychiatric scores predictions calculated over 50 runs. Vertical lines represent the standard deviations. The results marked fMRI, NEO, NEO+fMRI and NEO+fMRI+SNP are all based on the proposed model (different combinations of data). LRM is ridge regression, and LassoR is Lasso, and for both of these the NEO+fMRI+SNP data are all concatenated (all data are considered in these cases).

Figure 9 (left panel) shows a two-dimensional Isomap embedding representation of the learned factor scores (see Tenenbaum et al., 2000, for more details on Isomap). This 2D representation maintains geodesic distances between the subject scores as much as it is possible. In the plot, each point represents a subject, colored according to the cluster assignment. Note that there is a clear separation between the five groups. Cluster 1 is the biggest one with almost 30% of people in that group, whereas cluster 5 is the smallest with approximately 12% of people belonging to that group.

In attempt to draw broader inference regarding the nature of the five groups identified on the basis of NEO-PI-R responses, SNP data, fMRI responses and psychiatric scores, we examine the distribution of scores on internalizing, externalizing, and thought across these groups. Those scores are well-established measures which establish differences between individuals and provide risk for diverse psychiatric disorders. Thus, this exercise allows us to examine the extent to which our data-derived groups represent a unique window onto individual differences not captured by those standard measurements. Figure 9 (right panel) shows the average of the psychiatric scores across people within each group. Note that cluster 5 is characterized by people with high psychiatric disorder scores, which appears to be a high risk group for developing any form of psychopathology. Thus, our novel approach involving analysis of multi-view data appears positioned to identify individuals at greater general risk for psychiatric disorders (these results are consistent with those in Figure 8,

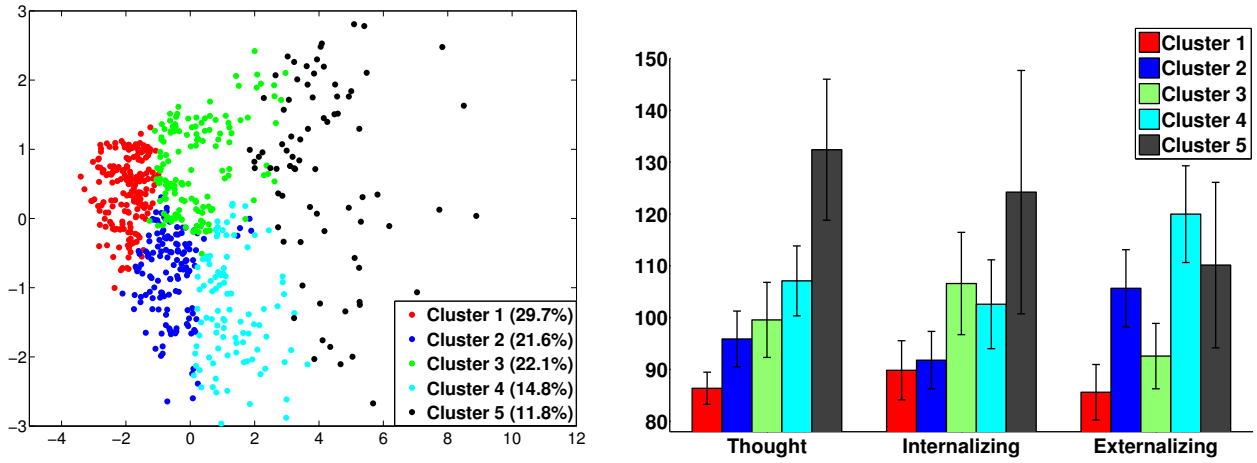


Figure 9: **Left**: Two-dimensional embedding of the factor scores, $\boldsymbol{V}$. Points are colored according to the cluster assignment. In parenthesis, the percentage of people belonging to each cluster. **Right**: Average of the psychiatric scores by cluster. Vertical lines represent the standard deviations.

where we accurately predicted the three psychiatric scores based on the available data).

6 Conclusions

In summary, using the proposed multi-modality factor analysis, we demonstrate the ability to accurately predict behaviorally and clinically relevant brain activity using information derived from self-report questionnaires and genetic data. A key component of our model is the development of a new framework for jointly learning from heterogeneous data. Given the relative ease with which self-report questionnaire data can be collected and modeled using the strategy introduced here, our findings hold promise for the identification of biomarkers for mental illness risk and treatment efficacy in routine clinical settings, in the absence of direct measures of brain activity through neuroimaging. This approach may prove similarly useful in population-based epidemiologic studies where neuroimaging may be impractical, by providing information on biomarkers representing mechanisms through which genes, environment, and their interactions affect disease status (Mitchell et al., 2011). Our Monte Carlo approach to testing, revealed that the prediction accuracy of our optimal model is both robust and stable. Nevertheless, it is of interest to extend tests of our predictive model to data from different tasks used to elicit brain activity, as well as data from different sample populations (e.g., adolescents, individuals at high-risk for disease).

References

Ahs, F., Davis, C., Gorka, A., and Hariri, A. (2014). Feature-based representations of emotional facial expressions in the human amygdala. *Social Cognitive and Affective Neuroscience*, 9:1372–1378.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London.

Bhattacharya, A. and Dunson, D. (2012). Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association*, 107:362–377.

Bickel, S. and Scheffer, T. (2004). Multi-View Clustering. In *Proceedings of the IEEE International Conference on Data Mining*.

Bryant, R. A., Felmingham, K., Kemp, A., Das, P., Hughes, G., Peduto, A., and Williams, L. (2008). Amygdala and ventral anterior cingulate activation predicts treatment response to cognitive behaviour therapy for post-traumatic stress disorder. *Psychol Med*, 38:555–561.

Carre, J., Hyde, L., Neumann, C., Viding, E., and Hariri, A. (2013). The neural signatures of distinct psychopathic traits. *Social Neuroscience*, 8:122–135.

Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103:1438–1456.

Casey, B. J., Ruberry, E. J., Libby, V., Glatt, C. E., Hare, T., Soliman, F., Duhoux, S., Frielingsdorf, H., and Tottenham, N. (2011). Transitional and translational studies of risk for anxiety. *Depress Anxiety*, 28:18–28.

Chaudhuri, K., Kakade, S. M., Livescu, K., and Sridharan, K. (2009). Multi-view Clustering via Canonical Correlation Analysis. In *Proceedings of the International Conference on Machine Learning*.

Chen, N., Zhu, J., and Xing, E. P. (2010). Predictive subspace learning for multiview data: a large margin approach. In *Advances in Neural Information Processing Systems*.

Cook, I. (2008). Biomarkers in psychiatry: Potentials, pitfalls, and pragmatics. *Primary Psychiatry*, 15:54–59.

Damianou, A. C., Ek, C. H., Titsias, M. K., and Lawrence, N. D. (2012). Manifold relevance determination. In *Proceedings of the International Conference on Machine Learning*.

Davis, M. and Whalen, P. (2001). The amygdala: vigilance and emotion. *Mol. Psychiatry*, 6:13–34.

Dunson, D. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104:1042–1051.

Elston, R. C. (2000). Introduction and overview. Statistical methods in genetic epidemiology. *Stat Methods Med Res*, 9(6):527–541.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.

Fakra, E., Hyde, L., Gorka, A., Fisher, P., Munoz, K., Kimak, M., Halder, I., Ferrell, R., Manuck, S., and Hariri, A. (2009). Effects of HTR1A C(-1019)G on amygdala reactivity and trait anxiety. *Arch. Gen. Psychiatry*, 66:33–40.

Forbes, E., Brown, S., Kimak, M., Ferrell, R., Manuck, S., and Hariri, A. (2009). Genetic variation in components of dopamine neurotransmission impacts ventral striatal reactivity associated with impulsivity. *Mol. Psychiatry*, 14:60–70.

Fyshe, A., Fox, E., Dunson, D., and Mitchell, T. (2012). Hierarchical latent dictionaries for models of brain activation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

Hariri, A. (2009a). The neurobiology of individual differences in complex behavioral traits. *Annu. Rev. Neurosci.*, 32:225–247.

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.

Hariri, A. R. (2009b). The neurobiology of individual differences in complex behavioral traits. *Annu Rev Neurosci*, 32:225–247.

Hariri, A. R., Brown, S. M., Williamson, D. E., Flory, J. D., de Wit, H., and Manuck, S. B. (2006). Preference for immediate over delayed rewards is associated with magnitude of ventral striatal activity. *Journal of Neuroscience*, 26(51):13213–13217.

Huettel, S., Song, A., and McCarthy, G. (2009). *Functional Magnetic Resonance Imaging*. Massachusetts: Sinauer.

Jia, Y., Salzmann, M., and Darrell, T. (2010). Factorized latent spaces with structured sparsity. In *Advances in Neural Information Processing Systems*.

Jovanovic, T. and Ressler, K. J. (2010). How the neurocircuitry and genetics of fear inhibition may inform our understanding of PTSD. *Am J Psychiatry*, 167:648–662.

Kareken, D., Claus, E., M., S., Dzemidzic, M., Kosobud, A., Radnovich, A., Hector, D., Ramchandani, V., O'Connor, S., Lowe, M., and Li, T. (2004). Alcohol-related olfactory cues activate the nucleus accumbens and ventral tegmental area in high-risk drinkers: preliminary findings. *Alcoholism: Clinical and Experimental Research*, 28(4):550–557.

Klami, A. and Kashi, S. (2007). Local dependent components. In *Proceedings of the International Conference on Machine Learning*.

Klami, A., Virtanen, S., and Kaski, S. (2013). Bayesian Canonical Correlation Analysis. *Journal of Machine Learning Research*, 14:965–1003.

Korgaonkar, M. S., Grieve, S. M., Etkin, A., Koslow, S. H., and Williams, L. M. (2013). Using standardized fMRI protocols to identify patterns of prefrontal circuit dysregulation that are common and specific to cognitive and emotional tasks in major depressive disorder: first wave results from the iSPOT-D study. *Neuropsychopharmacology*, 38(5):863–871.

Krueger, R. (1999). The structure of common mental disorders. *Archives of General Psychiatry*, 56:921–926.

Kumar, A., Rai, P., and Daumé III, H. (2011). Co-regularized Multi-view Spectral Clustering. In *Advances in Neural Information Processing Systems*.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

Luttinen, J. and Ilin, A. (2010). Transformations in variational Bayesian factor analysis to speed up learning. *Neurocomputing*, 73:1093–1102.

Mitchell, C., Notterman, D., Brooks-Gunn, J., Hobcraft, J., Garfinkel, I., Jaeger, K., Kotenko, I., and McLanahan, S. (2011). Role of mother's genes and environment in postpartum depression. *Proc Natl Acad Sci U S A*, 108:8189–8193.

Neil, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag.

Nikolova, Y., Bogdan, R., Brigidi, B., and Hariri, A. (2012). Ventral striatum reactivity to reward and recent life stress interact to predict positive affect. *Biol. Psychiatry*, 72:157–163.

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

Nikolova, Y. and Hariri, A. (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*, 2:19.

Nikolova, Y., Singhi, E., Drabant, E., and Hariri, A. (2013). Reward-related ventral striatum reactivity mediates gender-specific effects of a galanin remote enhancer haplotype on problem drinking. *Genes Brain Behav*, 12:516–524.

Polson, N. G. and Scott, J. G. (2010). Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction. In Bernardo, J., Bayarri, M., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics 9*, pages 76–106. Oxford University Press.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.

Qi, Y. and Jaakkola, T. (2007). Parameter expanded variational Bayesian methods. In *Advances in Neural Information Processing Systems*.

Salazar, E., Bogdan, R., Gorka, A., Hariri, A., and Carin, L. (2013). Exploring the mind: Integrating questionnaires and fMRI. In *Proceedings of the International Conference on Machine Learning*.

Salazar, E., Cain, M. S., Darling, E. F., Mitroff, S. R., and Carin, L. (2012). Inferring latent structure from mixed real and categorical relational data. In *Proceedings of the International Conference on Machine Learning*.

Schultz, R. (2005). Developmental deficits in social perception in autism: the role of the amygdala and fusiform face area. *Int. J. Dev. Neuroscience*, 23(2):125–141.

Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Buchel, C., ..., and Struve, M. (2010). The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. *Mol. Psychiatry*, 15:1128–1139.

Singh, I. and Rose, N. (2009). Biomarkers in psychiatry. *Nature*, 460:202–207.

Smith, K. (2012). Brain imaging: fMRI 2.0. *Nature*, 484:24–26.

Stein, J., Medland, S., Vasquez, A., Hibar, D., Senstad, R., Winkler, A., ..., and Muhleisen, T. (2012). Identification of common variants associated with human hippocampal and intracranial volumes. *Nat. Genet.*, 44:552–561.

Stingo, F., Guindani, M., Vannucci, M., and Calhoun, V. (2013). An integrative Bayesian modeling approach to imaging genetics. *Journal of the American Statistical Association*, 108:876–891.

Stoy, M., Schlagenhauf, F., Sterzer, P., Bermpohl, F., Hägele, C., Suchotzki, K., ..., and Ströhle, A. (2012). Hyporeactivity of ventral striatum towards incentive stimuli in unmedicated depressed patients normalizes after treatment with escitalopram. *Journal of Psychopharmacology*, 26(5):677–688.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 1:267–288.

Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244.

Virtanen, S., Klami, A., and Kaski, S. (2011). Bayesian CCA via group sparsity. In *Proceedings of the International Conference on Machine Learning*.

Virtanen, S., Klami, A., Khan, S. A., and Kaski, S. (2012). Bayesian group factor analysis. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

Wang, H., Nie, F., and Huang, H. (2013). Multi-view clustering and feature learning via structured sparsity. In *Proceedings of the International Conference on Machine Learning*.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics*, pages 723–732. Oxford University Press.

Whalen, P. J., Johnstone, T., Somerville, L. H., Nitschke, J. B., Polis, S., Alexander, A. L., Davidson, R. J., and Kalin, N. H. (2008). A functional magnetic resonance imaging predictor of treatment response to venlafaxine in generalized anxiety disorder. *Biol Psychiatry*, 63:858–863.

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.