Collaborative Phenotype Inference from Comorbid Substance Use Disorders and Genotypes

Jin Lu*, Jiangwen Sun*, Xinyu Wang*, Henry R. Kranzler†, Joel Gelernter‡ and Jinbo Bi*
*Department of Computer Science and Engineering, University of Connecticut, Storrs, Connecticut
†Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania
‡Departments of Psychiatry, Genetics, and Neurobiology, Yale University School of Medicine, New Haven, Connecticut
Email: jin.lu,jiangwen.sun,xinyu.wang,jinbo.bi@uconn.edu, kranzler@mail.med.upenn.edu, joel.gelernter@yale.edu

***Abstract*—Data in large-scale genetic studies of complex human diseases, such as substance use disorders, are often incomplete. Despite great progress in genotype imputation, e.g., the IMPUTE2 method, considerably less progress has been made in inferring phenotypes. We designed a novel approach to integrate individuals' comorbid conditions with their genotype data to infer missing (unreported) diagnostic criteria of a disorder. The premise of our approach derives from correlations among symptoms and the shared biological bases of concurrent disorders such as co-dependence on cocaine and opioids. We describe a matrix completion method to construct a bi-linear model based on the interactions of genotypes and known symptoms of related disorders to infer unknown values of another set of symptoms or phenotypes. An efficient stochastic and parallel algorithm based on the linearized alternating direction method of multipliers was developed to solve the proposed optimization problem. Empirical evaluation of the approach in comparison with other advanced data matrix completion methods via a case study shows that it both significantly improves imputation accuracy and provides greater computational efficiency.**

I. Introduction

Illicit drug use is very common in the United States and is associated with serious health and social problems [1]. Studies have shown that substance use disorders (SUDs) are heritable [2], [3], [4]. However, association studies aiming at identifying their genetic causes to date have been unsuccessful due to multiple factors, especially inadequate sample size [5]. In many datasets aggregated for genetic studies of mixed substance dependence, genetic data are often available for subjects exposed to one substance but not another [6]. These subjects are often excluded from the study because either they had no exposure to the substance or provided no reports of related symptoms [7], [6]. Dependence on different illicit drugs is correlated, both phenomenologically and biologically. Recent advances in the neurobiology of addiction have shown that many substances affect the same biological process, including reward, stress resiliency and executive cognitive control [8]. Many substance users use multiple drugs, resulting in comorbid dependence disorders [9], [10]. Sample size is essential to ensure adequate statistical power in genome-wide association studies (GWAS). The capability of expanding sample size by inferring missing phenotypes using the comorbidity among SUDs and shared genetic factors could be very helpful statistically. In this paper, we designed and evaluated such a statistical approach.

The problem of inferring diagnostic criteria of comorbid SUDs (CSUDs) can be considered to be analogous to a recommender system that predicts the preference of a user to a product with known preference for other products. We similarly would like to predict if a patient endorse a symptom based on the endorsement of other symptoms. By organizing the ratings of different users (rows) for various items (columns) into a matrix, a recommender system uses matrix completion methods to infer missing ratings. Similarly, by organizing symptoms of patients with a disorder into a matrix (as shown in Figure 1), we can use a matrix completion method to infer missing phenotypes. Classical matrix completion methods [11], [12] make use of the observation that the matrices to be completed are low rank (because similar users give similar ratings to similar products). Hence these methods search for solutions that minimize the rank of the completed matrix. Because similar patients may endorse similar symptoms, matrices that we seek to complete are expected to be low rank as well, so classic matrix completion methods may be applicable to our problem. However, these methods do not utilize additional useful information, such as associated genetic variants or known similarities between comorbid disorders, which is often referred to as side or auxiliary information in matrix completion. Therefore, classical methods may not be effective in solving our problem.

Recently matrix completion methods were proposed to make use of side information. Most of these methods are either non-convex [13], [14], or restricted in their use of side data to only in the minimization of matrix rank [15], [16], [17], leading to ambiguous or suboptimal solutions. A recent method completes a matrix $\mathbf{F}$ by considering the features ($\mathbf{X}$) describing row entities (e.g., users), side features ($\mathbf{Y}$) describing column entities (e.g., products), and their interactions ($\mathbf{X}^T\mathbf{G}\mathbf{Y}$) where $\mathbf{G}$ contains model parameters used in the inference model [18]. The optimization problem of this method is convex and the model parameter matrix $\mathbf{G}$ is not limited to being low rank. By a sampling rate of $O(logN)$, this approach could achieve exact recovery when the side features span the latent space of the matrix $\mathbf{F}$, and an ϵ-recovery when the side features are corrupted by perturbation. However, this method has limited scalability, and thus is difficult to be used in genetic studies with hundreds of symptoms or millions of genotypes. In this paper, we propose a novel stochastic

978-1-5090-3050-7/17/$31.00 ©2017 IEEE

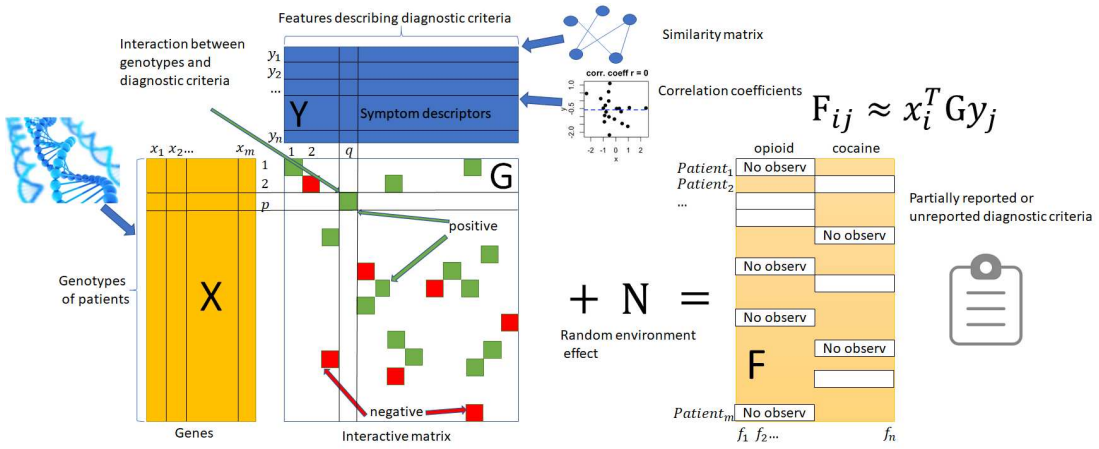

Fig. 1: Collaborative inference of CSUD diagnostic criteria using side information. Two sources including genotypes $\mathbf{X}$ and criterion similarity or correlation $\mathbf{Y}$ are integrated in a bi-linear model that predicts unreported diagnostic criteria. The matrix $\mathbf{G}$ captures the impact of the side features on the criteria to be inferred.

algorithm that is parallelized in a shared memory to solve the same optimization problem proposed in [18]. This algorithm can be proved to have global convergence and a sub-linear learning rate.

Figure 1 illustrates how we use our matrix completion method to infer missing phenotypes. Here, $\mathbf{F}$ is the phenotype matrix to be completed with rows representing patients and columns representing symptoms of a disorder, $\mathbf{X}$ contains genetic variants from different patients, $\mathbf{Y}$ characterizes similarities or correlations between each criterion of the disorder and criteria of other disorders. The inference model is in the form of $\mathbf{X}^T\mathbf{G}\mathbf{Y}+\mathbf{N}=\mathbf{F}$ where $\mathbf{N}$ captures any random effect from environment on the phenotype. The genetic variants in $\mathbf{X}$ are first identified through a GWAS. We use the proposed algorithm to determine $\mathbf{G}$ which is then used to infer missing phenotypes. We evaluated this approach in both simulations and the analysis of real world CSUD datasets and compared it with several state-of-the-art matrix completion methods.

The following notation is used throughout the paper. A vector is denoted by a bold lower case letter as in $\mathbf{v}$ and $\|\mathbf{v}\|_p$ represents its ℓ_p-norm that is defined by $\|\mathbf{v}\|_p = (|\mathbf{v}_{(1)}|^p + \cdots + |\mathbf{v}_{(d)}|^p)^{1/p}$, where $\mathbf{v}_{(i)}$ is the i-th index entry within the vector of $\mathbf{v}$ and d is the length of $\mathbf{v}$, also written as length$(\mathbf{v})$. A matrix is denoted by a bold upper case letter, e.g., $\mathbf{M}_{n\times d}$ is a n-by-d matrix, and $\|\mathbf{M}\|_F$ is its Frobenius norm.

II. Related Works

A recommender system, such as the Netflix movie recommendations, commonly uses collaborative filtering, or matrix completion, where the goal is to 'complete' the user-item rating matrix given a limited number of observed ratings. With the low-rank assumption of true underlying matrix, matrix completion methods require only the partially observed data in the matrix $\mathbf{F}$ and solve the following problem [11], [12]

$$\min_{\mathbf{E}} \|\mathbf{E}\|_*, \quad \text{subject to} \quad R_\Omega(\mathbf{E}) = R_\Omega(\mathbf{F}), \tag{1}$$

where $\mathbf{F} \in \mathbb{R}^{m\times n}$ is the partially observed low-rank matrix (with a rank of r) that needs to be recovered, $\Omega \subseteq \{1,\cdots,m\} \times \{1,\cdots,n\}$ be the set of indexes where the corresponding components in $\mathbf{F}$ are observed, the mapping $R_\Omega(\mathbf{M})$: $\mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ gives another matrix whose (i,j)-th entry is $\mathbf{M}_{i,j}$ if $(i,j) \in \Omega$ (or 0 otherwise), and $\|\mathbf{E}\|_*$ computes the nuclear norm of $\mathbf{E}$.

Several works utilize side features in their methods [13], [14] based on non-convex matrix factorization formulations with no theoretical guarantees. Three most recent methods have proposed convex formulations, which make it possible for them to have theoretical guarantees on matrix recovery [16], [17]. These methods all construct an inductive model $\mathbf{X}^T\mathbf{G}\mathbf{Y}$ so that $R_\Omega(\mathbf{X}^T\mathbf{G}\mathbf{Y}) = R_\Omega(\mathbf{F})$ where the side matrices $\mathbf{X}$ and $\mathbf{Y}$ consist of side features, respectively, for the row entities (e.g., users) and column entities (e.g., movies) of a (rating) matrix $\mathbf{F}$. This inductive model has a parameter matrix $\mathbf{G}$ that is either required to be sufficiently low rank [15] or to have a minimal nuclear norm $\|\mathbf{G}\|_*$ [16]. With a very strong assumption that both $\mathbf{X}$ and $\mathbf{Y}$ are orthonormal matrices, and respectively in the latent column and row space of the matrix $\mathbf{F}$, the method in [16] was proved to be likely to achieve an exact recovery of $\mathbf{F}$ with low sampling rate.

Another recent work [17] improves [16] by introducing a residual matrix $\mathbf{N}$ to handle the noisy side features. This method constructs an inductive model in the form of $\mathbf{X}^T\mathbf{G}\mathbf{Y}+\mathbf{N}$ to approximate $\mathbf{F}$ and requires both $\mathbf{G}$ and $\mathbf{N}$ to be low rank. An unnecessarily strong condition of the low-rank $\mathbf{G}$ is assumed because although a low-rank $\mathbf{G}$ leads to a low-rank $\mathbf{F}$, a low-rank $\mathbf{F}$ does not require a low-rank $\mathbf{G}$. Hence, we propose another method in [18] that removes this strong assumption. We will briefly revisit this approach in Section IV-B.

III. Data Description

A total of 7,189 subjects were aggregated from family and case-control based genetic studies of cocaine use disorder (CUD) and opioid use disorder (OUD). Subjects were recruited at five sites: Yale University School of Medicine, the University of Connecticut Health Center, the University of Pennsylvania Perelman School of Medicine, the Medical

TABLE I: Sample size by study and race

	African America	European America
CUD association, microarray	2,718	2,037
CUD association, exome sequencing	940	1,395
OUD association, microarray	1,398	1,756
OUD association, exome sequencing	540	1,190
Phenome inference	1,149	2,292

University of South Carolina, and McLean Hospital. The institutional review board at each site approved the study protocol and informed consent forms. The National Institute on Drug Abuse and the National Institute on Alcohol Abuse and Alcoholism each provided a Certificate of Confidentiality to protect participants. Subjects were paid for their participation. Among the total 7,189 subjects, 7,008 had cocaine exposure and were included in a GWAS of CUD and 4,843 had opioid exposure and were used in a GWAS of OUD. In total, 4,662 subjects had both exposures. Of that number, 3,441 subjects that had used an opioid more than 11 times were used in the evaluation of our approach to infer opioid use behaviors. Statistics for these datasets can be found in Table I.

Phenotypic information was obtained by face-to-face interview using the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA), a computer-assisted interview comprised of 26 sections (including sections for both cocaine and opioids) that yields diagnoses of various SUDs and Axis I psychiatric disorders, as well as antisocial personality disorder [19] For the DSM-5 diagnosis of CUD, OUD or SUD in general, 11 criteria, which can be clustered into four groups: impaired control, social impairment, risky use and pharmacological effects. The criteria for CUD and OUD were evaluated using questions from the SSADDA cocaine and opioid sections, respectively. In this study, we aimed to impute data for the 11 criteria for the drug that subjects had no prior exposure based on the criteria that they met for the use of other drugs. For example, we imputed cocaine use criteria from opioid use criteria, or vice-versa. In order to have groundtruth to evaluate the proposed and other compared methods, we include subjects for whom we have both cocaine and opioid (i.e., 3,441) symptoms.

Most of the sample subjects were genotyped using one of two different methods: Illumina HumanOmni1-Quad v1.0 microarray (MA) or exome sequencing (ES). There were total of 4,821 subjects genotyped with MA and 2,450 with ES. See [6] for details regarding the genotyping and variant calling. Genotypes were imputed with IMPUTE2 [20] using the genotyped variants and the 1000 Genomes reference panel (http://www.1000genomes.org/; released June 2011). For both the MA and ES sample, a total of 47,104,916 variants were imputed. We considered the imputed variants with $r^2 > 0.8$.

IV. Method

We describe the two steps in our proposed analysis where we first identified genetic variants in a GWAS and then used the identified genetic variants in a matrix completion method to complete a phenotype data matrix.

A. Finding genetic variants associated with CUD and OUD

The genetic relationship (GR) between each pair of subjects was evaluated with LDAK4 [21]. The evaluation was done separately in the MA and ES samples, and included only common variants with minor allele frequency (MAF) ≥ 0.03 and with a very high imputation quality (with $r^2 \geq 0.99$). There were 3,140,006 single nucleotide polymorphisms (SNPs) for MA and 604,884 for ES included in the GR estimation. The estimated GR matrix was used in subsequent association analyses to account for the population effect from genetic correlation.

To verify and correct the misclassification of self-reported race, we compared the MA (and ES) data from all subjects with genotypes from the HapMap 3 reference populations: CEU, YRI, and CHB. We conducted principal components (PC) analysis in the sample using PLINK [22] with 489,697 (91,089) SNPs common to those included in the GR evaluation in the MA (ES) dataset and HapMap panel [after pruning the MA (ES) SNPs for linkage disequilibrium (LD), defined as $r^2 > 80\%$] to characterize the underlying genetic architecture of the sample. The first PC scores distinguished African Americans (AAs) and European Americans (EAs), for which association analysis was done separately. The first three PCs were used in the analysis of each population to correct for residual population stratification.

The CUD (or OUD) criterion count was derived by counting the number of criteria endorsed out of the 11 DSM-5 criteria and was used in the GWAS to identify genetic variants. We ran the genomewide efficient mixed model association (GEMMA) method [23] to conduct association tests with sex and age, as well as the first three PCs being covariates. We combined the results from all eight studies (with the two different traits [CUD or OUD], datasets [MA or ES], or populations [AAs or EAs]) via meta analysis using METAL [24]. Genetic variants with meta P-value $< 1 \times 10^{-5}$ were used as side features in the subsequent phenotype inference.

B. Matrix completion with side information

We briefly review the formulation of the matrix completion method in [18]. This method integrates the side information into the formulation by explicitly building a bilinear predictive model that predicts missing components in the matrix ($\mathbf{F}$) using side features. Mathematically, we have $f = \mathbf{x}^T\mathbf{H}\mathbf{y} + \mathbf{x}^T\mathbf{u} + \mathbf{y}^T\mathbf{v} + g$, where $\mathbf{x}$ and $\mathbf{y}$ are the side feature vectors of a patient and a symptom, respectively, and $\mathbf{u}, \mathbf{v}, g$ and $\mathbf{H}$ are model parameters. By defining $\bar{\mathbf{x}} = [\mathbf{x}^T\ 1]^T$, $\bar{\mathbf{y}} = [\mathbf{y}^T\ 1]^T$ and $\mathbf{G}^{(a=d_1+1)\times(b=d_2+1)} = \begin{pmatrix} \mathbf{H} & \mathbf{u} \\ \mathbf{v}^T & g \end{pmatrix}$, the above equation can be simplified to: $f = \bar{\mathbf{x}}^T\mathbf{G}\bar{\mathbf{y}}$. We solve the following overall problem formulation for the best $\mathbf{G}$:

$$\begin{aligned} \min_{\mathbf{G},\mathbf{E}} \quad & \frac{1}{2}\|\mathbf{X}^T\mathbf{G}\mathbf{Y} - \mathbf{E}\|_F^2 + \lambda_E\|\mathbf{E}\|_* + \lambda_G g(\mathbf{G}), \\ \text{subject to} \quad & R_\Omega(\mathbf{E}) = R_\Omega(\mathbf{F}). \end{aligned} \tag{2}$$

where $\mathbf{E}$ is a completed version of $\mathbf{F}$. The $\mathbf{X}$ and $\mathbf{Y}$ here are matrices that are created by stacking one row of all ones to the

original $\mathbf{X}$ and $\mathbf{Y}$, respectively. To simplify the notation, we still used $\mathbf{X}$ and $\mathbf{Y}$ to represent the two augmented matrices. Because the phenotype data matrix is expected to be low rank, we also require $\mathbf{E}$ to be low rank, which is commonly translated into a minimization of the nuclear norm $\|\mathbf{E}\|_*$. Additionally, $g(\mathbf{G})$ is a function of $\mathbf{G}$ that imposes certain prior on $\mathbf{G}$. Because side features can be noisy and not all of them and their interactions are helpful in predicting $\mathbf{F}$, we expect $\mathbf{G}$ to be sparse and implement $g(\mathbf{G})$ with $\|\mathbf{G}\|_1$. The hyperparameters λ_E and λ_G help to balance the three components in the objective function and can be determined using cross validation.

Formulation (2) differs in several ways from existing methods that make use of side information for matrix completion. Besides the flexibility to consider both linear and quadratically interactive terms, this method allows the algorithm to determine the terms that should be used in the model by enforcing the sparsity in $\mathbf{G}$ instead of a non-sufficient condition to ensure low rank $\mathbf{E}$. Moreover, our formulation is still applicable when there is no access to useful side information by appropriately configuration of λ_G and λ_E. Related theoretical discussions can be found in [18].

V. Stochastic and parallel LADMM

In this section, we derive a stochastic version of the Linearized Alternating Direction Method of Multipliers (StoLADMM) from the classic LADMM algorithm [25] and further parallelize it to solve Problem (2). Besides the major advantage of computational efficiency, the algorithm gurantees the global convergence and a sub-linear learning rate.

To meet the condition that blocks of variables are separable, in order to use StoLADMM, we first defined $\mathbf{C} = \mathbf{E} - \mathbf{X}^T\mathbf{G}\mathbf{Y}$ and plugged it into Eq.(2). Following the LADMM scheme, the augmented Lagrangian function of (2) can be written as

$$\begin{aligned}&\mathcal{L}(\mathbf{E},\mathbf{G},\mathbf{C},\mathbf{M}_1,\mathbf{M}_2,\beta)\\ =&\frac{1}{2}\|\mathbf{C}\|_F^2 + \lambda_E\|\mathbf{E}\|_* + \lambda_G\|\mathbf{G}\|_1 + \frac{\beta}{2}\|R_\Omega(\mathbf{E}-\mathbf{F})\|_F^2\\ &+ \langle \mathbf{M}_1, R_\Omega(\mathbf{E}-\mathbf{F})\rangle + \left\langle \mathbf{M}_2, \mathbf{E}-\mathbf{X}^T\mathbf{G}\mathbf{Y}-\mathbf{C}\right\rangle\\ &+ \frac{\beta}{2}\|\mathbf{E}-\mathbf{X}^T\mathbf{G}\mathbf{Y}-\mathbf{C}\|_F^2\end{aligned}$$

where $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{m\times n}$ are called Lagrange multipliers and $\beta > 0$ is the penalty parameter. As a iterative algorithm, given $\mathbf{C}^k$, $\mathbf{G}^k$, $\mathbf{E}^k$, $\mathbf{M}_1^k$ and $\mathbf{M}_2^k$ at iteration k, we update each group of the variables while fixing others. The four steps are noted as Updating $\mathbf{C}$, Updating $\mathbf{E}$, Updating $\mathbf{G}$ and Updating $\mathbf{M}$ as below.

Updating $\mathbf{C}$:

$$\begin{aligned}\mathbf{C}^{k+1} =& \arg\min_{\mathbf{C}} \frac{1}{2}\|\mathbf{C}\|_F^2 + \left\langle \mathbf{M}_2^k, \mathbf{E}^k - \mathbf{X}^T\mathbf{G}^k\mathbf{Y} - \mathbf{C}\right\rangle\\ &+ \frac{\beta}{2}\|\mathbf{E}^k - \mathbf{X}^T\mathbf{G}^k\mathbf{Y} - \mathbf{C}\|_F^2\end{aligned}$$

which has a closed form solution as:

$$\mathbf{C}^{k+1} = \frac{\beta}{\beta+1}(\mathbf{E}^k - \mathbf{X}^T\mathbf{G}^k\mathbf{Y} + \mathbf{M}_2^k/\beta)$$

Updating $\mathbf{G}$:

$$\begin{aligned}&\min_{\mathbf{G}} \lambda_G\|\mathbf{G}\|_1 + \left\langle \mathbf{M}_2, \mathbf{E}^k - \mathbf{X}^T\mathbf{G}\mathbf{Y} - \mathbf{C}^k\right\rangle\\ &+ \frac{\beta}{2}\|\mathbf{E}^k - \mathbf{X}^T\mathbf{G}\mathbf{Y} - \mathbf{C}^k\|_F^2,\end{aligned} \tag{3}$$

after adding constant term to Eq. (3) we obtain

$$\min_{\mathbf{G}} \lambda_G\|\mathbf{G}\|_1 + \frac{\beta}{2}\|\mathbf{B}^k - \mathbf{X}^T\mathbf{G}\mathbf{Y}\|_F^2$$

where $\mathbf{B}_1^k = \mathbf{E}^k + \mathbf{M}_2^k/\beta - \mathbf{C}^k$. By converting the matrix $\mathbf{b}$ into a vector $\mathbf{g} = \text{vec}(\mathbf{G})$, $\text{vec}(\mathbf{X}^T\mathbf{G}\mathbf{Y}) = (\mathbf{Y}^T \otimes \mathbf{X}^T)\mathbf{g}$, further we let $\mathbf{b}^k = \text{vec}(\mathbf{B}^k)$ and $\otimes$ computes the Kronecker product of two matrices. Thus, if we denote $\mathbf{A} = (\mathbf{Y}^T \otimes \mathbf{X}^T)$, the above problem becomes:

$$\min_{\mathbf{g}} \lambda_G\|\mathbf{g}\|_1 + \frac{\beta}{2}\|\mathbf{A}\mathbf{g} - \mathbf{b}_1^k\|_2^2 \tag{4}$$

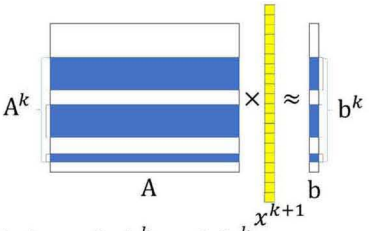

Fig. 2: Calculation of $\mathbf{A}^k$ and $\mathbf{b}^k$. s rows are randomly sampled from $\mathbf{A}$ and $\mathbf{b}$ individually.

Here (4) is a lasso problem, which has to be solved iteratively in practice, making it problematic to compute or even memorize when the size of matrix $\mathbf{A}$ becomes extremely large, as occurs in many real cases. By utilizing the stochasticity and linearization technique in ADMM, we approximate our problem as

$$\|\mathbf{A}^k\mathbf{g} - \tilde{\mathbf{b}_1^k}\|_2^2 \approx \|\mathbf{A}^k\mathbf{g}^k - \tilde{\mathbf{b}_1^k}\|_2^2 + 2\left\langle f_1^k, \mathbf{g}-\mathbf{g}^k\right\rangle + \tau_A\|\mathbf{g}-\mathbf{g}^k\|_2^2$$

where $\tilde{\mathbf{A}^k}$ and $\tilde{\mathbf{b}^k}$ sample at random from the corresponding s rows of $\mathbf{A}$ and $\mathbf{b}^k$ in pairs, as shown in Figure 2. $\tau_A > 0$ is a proximal parameter and

$$f_1^k = \tilde{\mathbf{A}^k}^T(\tilde{\mathbf{A}^k}\mathbf{g}^k - \tilde{\mathbf{b}_1^k}) \tag{5}$$

is the stochastic gradient of $\frac{1}{2}\|\mathbf{A}\mathbf{g} - \mathbf{b}_1^k\|_2^2$ at $\mathbf{g}_k$. The stochastic approximation can tremendously reduce memory consumption and save computational costs in each iteration. Then the above equation can be re-written as:

$$\min_{\mathbf{g}} \lambda_G\|\mathbf{g}\|_1 + \frac{\beta\tau_A}{2}\|\mathbf{g} - [\mathbf{g}^k - f_1^k/\tau_A]\|_2^2$$

Obviously the closed-form solution is:

$$\mathbf{g}^{k+1} = \max(|\mathbf{g}^k - f_1^k/\tau_A| - \frac{\lambda_G}{\tau_A\beta}, 0) \odot sgn(\mathbf{g}^k - f_1^k/\tau_A)$$

In advance, our efficient procedure calculates each stochastic gradient in parallel by using multiple computational nodes, i.e., workers, then averaging those gradient values by a central computational node, i.e., a master.

Updating $\mathbf{E}$:

$$\begin{aligned}&\min_{\mathbf{E}} \lambda_E\|\mathbf{E}\|_* + \left\langle \mathbf{M}_1^k, R_\Omega(\mathbf{E}-\mathbf{F})\right\rangle + \frac{\beta}{2}\|R_\Omega(\mathbf{E}-\mathbf{F})\|_F^2\\ &+ \left\langle \mathbf{M}_2^k, \mathbf{E} - \mathbf{X}^T\mathbf{G}^{k+1}\mathbf{Y} - \mathbf{C}^k\right\rangle + \frac{\beta}{2}\|\mathbf{E} - \mathbf{X}^T\mathbf{G}^{k+1}\mathbf{Y} - \mathbf{C}^k\|_F^2\end{aligned}$$

where we can re-organize this subproblem in a simpler form as:

$$\min_{\mathbf{E}} \lambda_E \|\mathbf{E}\|_* + \frac{\beta}{2}\|R_\Omega(\mathbf{E} - \mathbf{B}_2^k)\|_F^2 + \frac{\beta}{2}\|\mathbf{E} - \mathbf{B}_3^k\|_F^2$$

where $\mathbf{B}_2^k = R_\Omega(\mathbf{F} - \mathbf{M}_1^k/\beta)$ and $\mathbf{B}_3^k = \mathbf{X}^T\mathbf{G}^{k+1}\mathbf{Y} + \mathbf{C}^k - \mathbf{M}_2^k/\beta$. Via the same linearization technique, the problem can be approximated by:

$$\min_{\mathbf{E}} \lambda_E \|\mathbf{E}\|_* + \frac{\beta\tau_B}{2}\|\mathbf{E} - (\mathbf{E}^k - f_2^k/\tau_B)\|_F^2 + \frac{\beta\tau_B}{2}\|\mathbf{E} - (\mathbf{E}^k - f_3^k/\tau_B)\|_F^2$$

where f_2^k and f_3^k are the gradients of $\frac{1}{2}\|R_\Omega(\mathbf{E} - \mathbf{B}_2^k)\|_F^2$ and $\frac{1}{2}\|\mathbf{E} - \mathbf{B}_3^k\|_F^2$ at $\mathbf{E}^k$, which are illustrated below:

$$\begin{aligned} f_2^k &= R_\Omega(\mathbf{E}^k - \mathbf{B}_2^k) = R_\Omega(\mathbf{E}^k - \mathbf{F} + \mathbf{M}_1^k/\beta), \\ f_3^k &= \mathbf{E}^k - \mathbf{B}_3^k = \mathbf{E}^k - \mathbf{X}^T\mathbf{G}^{k+1}\mathbf{Y} - \mathbf{C}^k + \mathbf{M}_2^k/\beta. \end{aligned} \tag{6}$$

Therefore, the closed-form solution can be obtained as

$$\mathbf{E}^{k+1} = SVT(\mathbf{E}^k - (f_2^k + f_3^k)/(2\tau_B), \lambda_E/2(\beta\tau_B))$$

Here the operator $SVT(\mathbf{E}, t)$ is defined in [11] for soft-thresholding the singular values of an arbitrary matrix $\mathbf{E}$ by t.

Updating $\mathbf{M}$:

$$\begin{aligned} \mathbf{M}_1^{k+1} &= \mathbf{M}_1^k + \beta(R_\Omega(\mathbf{E}^{k+1} - \mathbf{F})), \\ \mathbf{M}_2^{k+1} &= \mathbf{M}_2^k + \beta(\mathbf{E}^{k+1} - \mathbf{X}^T\mathbf{G}^{k+1}\mathbf{Y} - \mathbf{C}^{k+1}). \end{aligned}$$

Algorithm 1 summarizes the StoLADMM steps for the variables of $(\mathbf{C}, \mathbf{E}, \mathbf{G}, \mathbf{M}_1, \mathbf{M}_2)$.

Algorithm 1 The StoLADMM algorithm to solve $\mathbf{C}^k$, $\mathbf{G}^k$, $\mathbf{E}^k$, $k = 1, ..., K$

Input: $\mathbf{X}$, $\mathbf{Y}$ and $R_\Omega(\mathbf{F})$ with parameters λ_G, λ_E, τ_A, τ_B, ρ and β_{max}.

Output: $\mathbf{C}, \mathbf{G}, \mathbf{E}$;

Initialize $\mathbf{E}^0, \mathbf{G}^0, \mathbf{M}_1^0, \mathbf{M}_2^0$. Compute $\mathbf{A} = \mathbf{Y}^T \otimes \mathbf{X}^T$. $k = 0$,

repeat;

$\mathbf{C}^{k+1} = \frac{\beta}{\beta+1}(\mathbf{E}^k - \mathbf{X}^T\mathbf{G}^k\mathbf{Y} + \mathbf{M}_2^k/\beta)$;

$\mathbf{G}^{k+1} = reshape(\max(|\mathbf{g}^k - f_1^k/\tau_A| - \frac{\lambda_G}{\tau_A\beta}, 0) \odot sgn(\mathbf{g}^k - f_1^k/\tau_A))$ where f_1^k can be computed by (5);

$\mathbf{E}^{k+1} = SVT(\mathbf{E}^k - (f_2^k + f_3^k)/(2\tau_B), \lambda_E/2(\beta\tau_B))$ where f_2^k and f_3^k can be computed by (6);

$\mathbf{M}_1^{k+1} = \mathbf{M}_1^k + \beta(R_\Omega(\mathbf{E}^{k+1} - \mathbf{F}))$.

$\mathbf{M}_2^{k+1} = \mathbf{M}_2^k + \beta(\mathbf{E}^{k+1} - \mathbf{X}^T\mathbf{G}^{k+1}\mathbf{Y} - \mathbf{C}^{k+1})$.

$k = k + 1$ until convergence;

Return $\mathbf{C}, \mathbf{G}, \mathbf{E}$;

It is beneficial that our algorithm has an $O(1/\sqrt{t})$ convergence rate, the same as the convergence rate as proved in [26], which guarantees our algorithm's performance while saving considerable memory and computational costs. In Algorithm 1 we set the sampling block size s to $\max(1, \sqrt{\text{length}(\mathbf{g})/100})$. $\tau_A < \|\mathbf{A}\|$, $\tau_B < \|R_\Omega(\mathbf{F})\|$ and $\beta = 0.01$ as the preferable values [25] in practice. In the initialization step, $\mathbf{M}_1^0$ and $\mathbf{M}_2^0$ are randomly drawn from the standard Gaussian distribution; we initialize $\mathbf{E}_0$ and $\mathbf{G}_0$ by the iterative soft-thresholding algorithm [27] and SVT operator respectively.

q		StoLADMM	LADMM	DirtyIMC	IMC	MAXIDE	BM
20%	RMSE	0.236	0.231	0.297	**0.230**	0.235	0.567
	time(s)	30.938	664.515	45.366	**21.053**	4732.718	NaN
40%	RMSE	**0.226**	0.234	0.298	0.235	0.236	0.582
	time(s)	29.953	982.212	21.063	**20.803**	3772.202	NaN
60%	RMSE	**0.228**	0.236	0.301	0.237	0.235	0.581
	time(s)	28.719	815.841	**20.269**	36.737	4718.916	NaN
80%	RMSE	**0.236**	0.237	0.303	0.239	0.241	0.585
	time(s)	30.547	877.886	**23.906**	32.872	4011.692	NaN
100%	RMSE	**0.223**	0.239	0.303	0.246	0.242	0.574
	time(s)	30.172	489.770	**22.922**	24.653	3695.292	NaN

TABLE II: The inference results on the Opioid-Cocaine data.

VI. Evaluation

We compared the proposed method with other matrix completion approaches that also use side information, including: MAXIDE [16], IMC [15] and DirtyIMC [17]. A benchmark method (BM) was also compared, which simply used the known values on the comparable phenotypes between cocaine and opioid use disorders to impute missing entries. We randomly removed the phenotypes of q% CSUD patients associated with either opioid or cocaine use (not both). Then, all compared methods were run to infer these missing values. Their performance was measured by the relative mean squared error (RMSE) calculated on missing entries: $\|R_{\Omega^c}(\mathbf{X}^T\mathbf{G}\mathbf{Y} - \mathbf{F})\|_2^2/\|R_{\Omega^c}(\mathbf{F})\|_2^2$. The hyperparameters λ's and the rank of $\mathbf{G}$ (required by IMC and DirtyIMC) were selected by cross-validation: using 30% of the given entries randomly to form a validation set. Models were obtained by applying each method to the remaining entries with a specific choice of λ from $10^{-3}, 10^{-2}, ..., 10^4$. The average validation RMSE was computed by repeating the above procedure three times. The hyperparameter values were fixed to those that gave the best average validation RMSEs for each individual method. For each choice of q, we repeated the above entire procedure five times and reported the average RMSE on the missing entries. All tests were conducted using Matlab on an Intel Core i7 3.6GHz and 16GB RAM computer.

The 383 genetic variants identified in our GWAS were used as side features $\mathbf{X}$. The correlations between 22 CUD and 22 OUD symptoms formed a correlation matrix which was used as side features $\mathbf{Y}$. The accuracy and computational time for all methods with five q values (ranging from 20% to 100%) are presented in Table II. The results indicate that our method improves computational efficiency without sacrificing recovery accuracy. When the run time was significantly reduced by nearly 95% of that used by the non-stochastic LADMM, our RMSE still outperformed other methods, showing that our method can readily handle big data.

Figure 3 depicts the recovered $\mathbf{G}$ matrix whose rows (corresponding to genotypes) are sorted in ascending order according to their association p-values, and columns correspond to 22 phenotypes. Color of higher saturation reflects the stronger interaction between a specific genotype and a criterion. Red represents a positive interaction, while blue represents a negative interaction. We observed that the top 20 rows had the most none zero values in the matrix because these rows corresponded to the most significantly associated genetic variants. For instance, the interactions between a

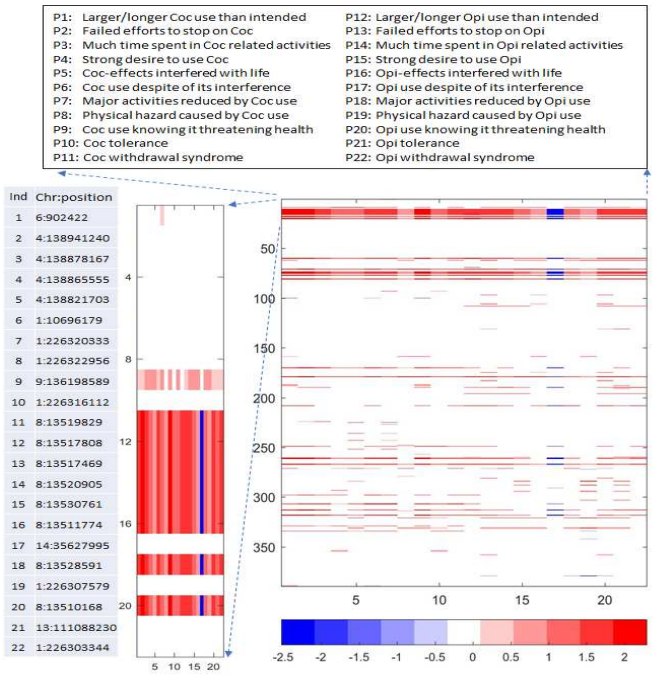

Fig. 3: Recovered $\mathbf{G}$ by our method for the CSUD dataset.

marker at chrome 9 and position 136198589 and all the 22 phenotypes had the largest weights in the inference model for the missing criteria. This shows that our phenotype inference framework can not only use additive effects of genotypes but also interactive effects between genotype and phenotype.

VII. Conclusion

We adapted a matrix completion approach to infer SUD criteria using both correlation among criteria of different conditions and genotypes as side information. By imposing sparse prior on the model parameters, the method can find a sparse interactive matrix that connects specific genotypes to diagnostic criteria. We introduced an efficient stochastic LADMM algorithm to solve the optimization problem in this method. The empirical evaluation shows that our method can significantly enhance the running efficiency with minimal adverse effects on the imputation accuracy.

Acknowledgment

Correspondence should be addressed to Jinbo Bi. This work was supported by NSF grants CCF-1514357 and DBI-1356655, and NIH grants R01DA037349 and K02DA043063.

References

[1] Center for Behavioral Health Statistics and Quality, "Key Substance Use and Mental Health Indicators in the United States: Results from the 2015 National Survey on Drug Use and Health," *HHS Publication No. SMA 16-4984, NSDUH Series H-51*, 2016.

[2] K. S. Kendler, L. M. Karkowski, M. C. Neale, and C. A. Prescott, "Illicit psychoactive substance use, heavy use, abuse, and dependence in a US population-based sample of male twins," *Arch Gen Psychiatry*, no. 3, pp. 261–169.

[3] J. Bi, J. Gelernter, J. Sun, and H. R. Kranzler, "Comparing the utility of homogeneous subtypes of cocaine use and related behaviors with DSM-IV cocaine dependence as traits for genetic association analysis," *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, vol. 165, no. 2, pp. 148–156, 2014.

[4] J. Sun, J. Bi, G. Chan, D. Oslin, L. Farrer, J. Gelernter, and H. R. Kranzler, "Improved methods to identify stable, highly heritable subtypes of opioid use and related behaviors," *Addictive Behaviors*, vol. 37, no. 10, pp. 1138–1144, 2012.

[5] K. P. Jensen, "A Review of Genome-Wide Association Studies of Stimulant and Opioid Use Disorders." *Molecular neuropsychiatry*, no. 1, pp. 37–45.

[6] J. Gelernter, H. R. Kranzler, R. Sherva, R. Koesterer, L. Almasy, H. Zhao, and L. a. Farrer, "Genome-wide association study of opioid dependence: Multiple associations mapped to calcium and potassium pathways," *Biological Psychiatry*, vol. 76, pp. 66–74, 2014.

[7] J. Sun, H. R. Kranzler, and J. Bi, "An Effective Method to Identify Heritable Components from Multivariate Phenotypes," *PLoS ONE*, vol. 10, no. 12, pp. 1–22, 2015.

[8] D. Goldman, G. Oroszi, and F. Ducci, "The genetics of addictions: uncovering the genes." *Nature reviews. Genetics*, no. 7, pp. 521–32.

[9] R. Hammersley, A. Forsyth, and T. Lavelle, "The criminality of new drug users in glasgow," *Addiction*, vol. 85, no. 12, pp. 1583–1594, 1990.

[10] J. C. Ball and A. Ross, *The effectiveness of methadone maintenance treatment: patients, programs, services, and outcome*. Springer Science & Business Media, 2012.

[11] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. on Optimization*, vol. 20, no. 4, pp. 1956–1982, Mar. 2010.

[12] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *Information Theory, IEEE Transactions on*, vol. 56, no. 6, pp. 2980–2998, June 2010.

[13] A. K. Menon, K.-P. Chitrapura, S. Garg, D. Agarwal, and N. Kota, "Response prediction using collaborative filtering with hierarchies and side-information," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 141–149.

[14] N. Natarajan and I. S. Dhillon, "Inductive matrix completion for predicting gene–disease associations," *Bioinformatics*, vol. 30, no. 12, pp. i60–i68, 2014.

[15] P. Jain and I. S. Dhillon, "Provable inductive matrix completion," *arXiv preprint arXiv:1306.0626*, 2013.

[16] M. Xu, R. Jin, and Z. hua Zhou, "Speedup matrix completion with side information: Application to multi-label learning," *NIPS*, pp. 2301–2309, 2013.

[17] K.-Y. Chiang, C.-J. Hsieh, and I. S. Dhillon, "Matrix completion with noisy side information," *NIPS*, pp. 3429–3437, 2015.

[18] J. Lu, G. Liang, J. Sun, and J. Bi, "A sparse interactive model for matrix completion with side information," in *NIPS*, 2016, pp. 4071–4079.

[19] A. Pierucci-Lagha, J. Gelernter, R. Feinn, J. F. Cubells, D. Pearson, A. Pollastri, L. Farrer, and H. R. Kranzler, "Diagnostic reliability of the semi-structured assessment for drug dependence and alcoholism (SSADDA)," *Drug and Alcohol Dependence*, vol. 80, no. 3, pp. 303–312, 2005.

[20] B. N. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS Genetics*, vol. 5, no. 6, 2009.

[21] D. Speed, N. Cai, M. Johnson, S. Nejentsev, and D. Balding, "Re-evaluation of SNP heritability in complex human traits," *bioRxiv*, no. 7, p. 074310.

[22] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: A tool set for whole-genome association and population-based linkage analyses," *American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[23] X. Zhou and M. Stephens, "Genome-wide efficient mixed-model analysis for association studies," *Nature genetics*, no. 7, pp. 821–4.

[24] C. J. Willer, Y. Li, and G. R. Abecasis, "METAL: Fast and efficient meta-analysis of genomewide association scans," *Bioinformatics*, vol. 26, no. 17, pp. 2190–2191, 2010.

[25] J. Yang and X.-M. Yuan, "Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization," *Math. Comput.*, vol. 82, 2013.

[26] H. Ouyang, N. He, L. Tran, and A. Gray, "Stochastic alternating direction method of multipliers," in *International Conference on Machine Learning*, 2013, pp. 80–88.

[27] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.