IMAGING GENETICS VIA SPARSE CANONICAL CORRELATION ANALYSIS

Eric C. Chi[⋆]*, Genevera I. Allen*[∗]*, Hua Zhou*[‡]*, Omid Kohannim*[†]*, Kenneth Lange*[⋆]*, Paul M. Thompson*[†]

[⋆] Department of Human Genetics, UCLA School of Medicine, Los Angeles, CA, USA
[∗] Department of Statistics, Rice University, Houston, TX, USA
[‡] Department of Statistics, North Carolina State University, Raleigh, NC, USA
[†] Imaging Genetics Center, Lab. of Neuro Imaging, UCLA School of Medicine, Los Angeles, CA, USA

ABSTRACT

The collection of brain images from populations of subjects who have been genotyped with genome-wide scans makes it feasible to search for genetic effects on the brain. Even so, multivariate methods are sorely needed that can search both images and the genome for relationships, making use of the correlation structure of both datasets. Here we investigate the use of sparse canonical correlation analysis (CCA) to home in on sets of genetic variants that explain variance in a set of images. We extend recent work on penalized matrix decomposition to account for the correlations in both datasets. Such methods show promise in imaging genetics as they exploit the natural covariance in the datasets. They also avoid an astronomically heavy statistical correction for searching the whole genome and the entire image for promising associations.

***Index Terms*—** Diffusion tensor imaging, Genome wide association, Canonical correlation analysis, sparsity, lasso

1. INTRODUCTION

The last few years have seen an unprecedented surge in data acquisition in fields ranging from signal processing to biology and medicine. This ability to acquire massive amounts of data has opened the door to qualitatively different approaches to science as well, often using high-dimensional datasets from more than one modality. One example is the convergence of biomedical imaging and genomics in the nascent field of imaging genetics [1]. The basic idea is to identify genetic variants that can best capture and explain phenotypic variations in brain function and structure. To be more concrete, two sets of data are observed, p genotypes and q neuroimaging phenotypes, on n samples. Both p and q may be small or large and there has been prior testing for effects in various scenarios. In [2], Joyner et al. studied a dataset with small q - four brain size measures -and small p - 11 SNPs. In [3], Potkin et al. considered small q, the mean BOLD signal from fMRI, and large p, 317,503 SNPs. Filippini et al. explored the combination of large q, 29,812 voxels, and small p, a single SNP [4]. Finally, Stein et al. in [5] took on the most challenging scenario of large q, 31,622 voxels, and large p, 448,293 SNPs. Thus, on the same set of imaged subjects, high-dimensional genetic data is also collected, e.g. hundreds of thousands of SNP genotypes. In some cases, well defined regions of interest (ROIs) are already known, but in other cases they are not. Similarly, in some cases, candidate genes may or may not be available. In this scenario, one wishes to simultaneously identify ROIs and a parsimonious set of genetic loci that are associated with each other.

The last case in particular presents not only intriguing possibilities but also computational and statistical challenges. Indeed, the simplest strategy is to perform pq univariate regressions between all possible voxels and SNPs [5] and adjust for multiple comparisons. While such an approach is straightforward, it also completely ignores the correlation structure among the SNPs and voxels. It also lacks power, as an astronomical correction must be made for the number of tests performed. Given the correlation structure a multivariate approach is called for. To address these shortcomings, in this paper we present a sparse canonical correlation analysis (CCA) method to identify joint signals in a pair of high-dimensional data sets, namely diffusion tensor images (DTI) and single nucleotide polymorphism (SNP) measurements. The goal in classical CCA is to determine a coordinate system that maximizes the cross-covariance between two data sets [6]. In other words we seek the linear transformation of two data sets such that the linear forms are maximally correlated. We anticipate, however, that relatively few DTI voxels will contain signals that are correlated with again relatively few SNPs. Despite the fact that both data live in very high dimensional spaces, DTI voxels number in the tens of thousands and SNPs number in the hundreds of thousands, the relevant signal often resides in a low dimensional manifold. Indeed penalized methods such as the LASSO [7] have been very successful in recovering meaningful parsimonious models from high dimensional data. Building on this idea, Witten et al. introduced a penalized matrix decomposition (PMD) on the sample cross-covariance matrix in [8] aimed at introducing sparsity into the linear combinations. In related work, Vounou et al. introduced sparse Reduced Rank Regression (sRRR) [9, 10] as another multivariate alternative. Indeed, the PMD is a special case of

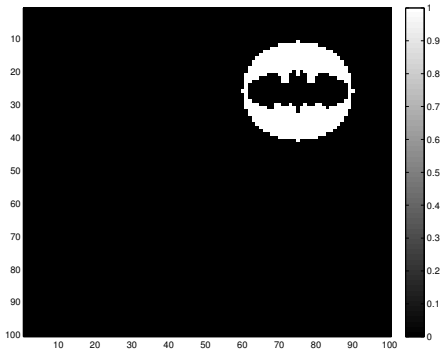

Fig. 1. Region of Interest (ROI) for a simulated example problem, with a coherent signal in the image. Pixels that belong to the ROI have value of 1. Pixels outside the ROI have value of 0.

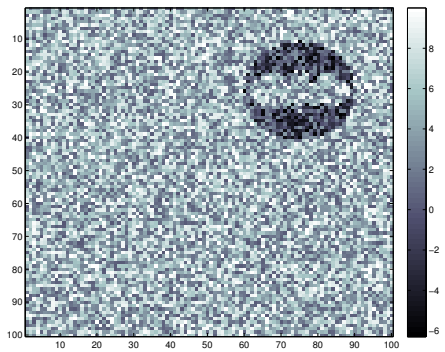

Fig. 2. An example of the observed image data: $\mathcal{Z}(1,:,:)$ which corresponds to the observed image of the first subject.

the sRRR when relevant covariance matrices are taken to be identity matrices. Nonetheless, despite having a more general framework, the algorithms presented in [9, 10] make the same simplifying assumptions made in PMD. In this work we extend the PMD model to account for correlation structure in both data sets.

2. METHODS

2.1. Simulation Experiment

Suppose we have 100 subjects for whom 1,000 SNPs have been typed, $\mathbf{X} \in \mathbb{R}^{100\times 1,000}$, and on which a 100×100 image has been taken, $\mathcal{Y} \in \mathbb{R}^{100\times 100\times 100}$ (here we introduce the ideas for a set of 2D images but they hold in any dimension without loss of generality). Let $\mathbf{Y} \in \mathbb{R}^{100\times 10,000}$ denote the matricization of $\mathcal{Y}$ along its first mode. The data sets are generated as follows. Let $\boldsymbol{\beta} \in \mathbb{R}^{1,000}$ be sparse with

$$\beta_i = \begin{cases} 2 & i \in \{192, 438, 623, 786, 780\} \\ 0 & \text{otherwise.} \end{cases}$$

The images are also similarly sparse.

$$y_{\mathbf{ij}} = \begin{cases} \sum_{j'} x_{ij'}\beta_{j'} & (i,j) \in R \subset \{1,\ldots,100\}^2 \\ 0 & \text{otherwise,} \end{cases}$$

where R denotes the ROI. The ROI for this problem is shown in Figure 1. Finally, we do not observe $\mathcal{Y}$ but rather a noisy version of it $\mathcal{Z}$.

$$z_{\mathbf{ijk}} = y_{\mathbf{ijk}} + \sigma\varepsilon_{\mathbf{ijk}},$$

where $\varepsilon_{\mathbf{ijk}}$ are i.i.d. standard normal and $\sigma = 10$. An example of the observed image data for a subject is shown in Figure 2.

2.2. Sparse Canonical Correlation Analysis

Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ denote the SNP data matrix and $\mathbf{Y} \in \mathbb{R}^{n\times q}$ denote the matrix of vectorized DTI fractional anisotropy (FA) scores. Classical CCA solves the following optimization problem

$$\max_{\mathbf{a},\mathbf{b}} \mathbf{a}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{Y}\mathbf{b}$$

subject to the constraints $\mathbf{a}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{a} = 1$ and $\mathbf{b}^\mathsf{T}\mathbf{Y}^\mathsf{T}\mathbf{Y}\mathbf{b} = 1$. The matrices $\mathbf{X}^\mathsf{T}\mathbf{Y}, \mathbf{X}^\mathsf{T}\mathbf{X}$, and $\mathbf{Y}^\mathsf{T}\mathbf{Y}$ are estimates of the cross-covariance and covariance matrices respectively. PMD introduces a LASSO penalty and assumes the covariance matrices are identity matrices, i.e., PMD solves the optimization problem

$$\max_{\mathbf{a},\mathbf{b}} \mathbf{a}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{Y}\mathbf{b} - \lambda_a\|\mathbf{a}\|_1 - \lambda_b\|\mathbf{b}\|_1$$

subject to the constraints $\mathbf{a}^\mathsf{T}\mathbf{a} \leq 1$ and $\mathbf{b}^\mathsf{T}\mathbf{b} \leq 1$. Note that the equality constraints have been relaxed to inequality constraints to make the feasible sets convex. The parameters $\lambda_a \geq 0$ and $\lambda_b \geq 0$ tune the degree of sparsity in $\mathbf{a}$ and $\mathbf{b}$. Since the objective function is biconvex, namely it is convex in $\mathbf{a}$ with $\mathbf{b}$ fixed and vice versa, PMD iteratively minimizes with respect to $\mathbf{a}$ holding $\mathbf{b}$ fixed, and vice versa until convergence. The update for $\mathbf{a}$ is given by

$$\begin{aligned} \hat{\mathbf{a}} &= \arg\min_{\mathbf{a}} \frac{1}{2}\|\mathbf{X}^\mathsf{T}\mathbf{Y}\mathbf{b} - \mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1 \\ \mathbf{a}^* &= \begin{cases} \frac{\hat{\mathbf{a}}}{\|\hat{\mathbf{a}}\|_2} & \text{if } \|\hat{\mathbf{a}}\|_2 > 0. \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The update for $\mathbf{b}$ is similar. To weaken the identity covariance assumption we minimize the same objective function but alter the constraints to $\mathbf{a}^\mathsf{T}\tilde{\boldsymbol{\Sigma}}_x\mathbf{a} \leq 1$ and $\mathbf{b}^\mathsf{T}\tilde{\boldsymbol{\Sigma}}_y\mathbf{b} \leq 1$, where $\tilde{\boldsymbol{\Sigma}}_x$ and $\tilde{\boldsymbol{\Sigma}}_y$ are estimated covariance matrices. Again the problem is amenable to block relaxation, namely iteratively minimizing with respect to $\mathbf{a}$ holding $\mathbf{b}$ fixed and vice versa. Consider

optimizing with respect to $\mathbf{a}$ first. We can rewrite the problem as

$$\max \mathbf{a}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{Y}\mathbf{b} - \lambda\|\mathbf{a}\|_1$$

subject $\mathbf{a}^\mathsf{T}\tilde{\mathbf{\Sigma}}_x\mathbf{a} \leq 1$. A little convex calculus shows that the updates are given by

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \frac{1}{2}\|\tilde{\mathbf{\Sigma}}_x^{-1/2}\mathbf{X}^\mathsf{T}\mathbf{Y}\mathbf{b} - \tilde{\mathbf{\Sigma}}_x^{1/2}\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1. \quad (1)$$

and

$$\mathbf{a}^* = \begin{cases} \frac{\hat{\mathbf{a}}}{\|\tilde{\mathbf{\Sigma}}_x\hat{\mathbf{a}}\|_2} & \text{if } \|\tilde{\mathbf{\Sigma}}_x\hat{\mathbf{a}}\|_2 > 0. \\ \mathbf{0} & \text{o.w.} \end{cases} \quad (2)$$

Thus, the update occurs in two stages. We first solve a LASSO penalized regression problem in (1) where the response variable is $\tilde{\mathbf{\Sigma}}_x^{-1/2}\mathbf{X}^\mathsf{T}\mathbf{Y}\mathbf{b}$ and the design matrix is $\tilde{\mathbf{\Sigma}}_x^{1/2}$. Then if the solution of (1) is non-zero we normalize the solution so that $\|\tilde{\mathbf{\Sigma}}_x\mathbf{a}^*\|_2 = 1$. If the solution to (1) is zero, the final solution $\mathbf{a}^*$ is zero. Note that if we take $\tilde{\mathbf{\Sigma}}_x$ and $\tilde{\mathbf{\Sigma}}_y$ to be identity matrices, we recover the algorithm employed in prior work [8, 9, 10].

We note that the choice of covariance estimator is critical. Indeed the sample covariance is well recognized as a poor estimator of the population covariance in the small n, large p regime considered here. This problem even plagues the classical CCA problem as well when p is close to n. This has been addressed by applying a ridge estimate of the covariance matrices, namely [11, 12]. Ledoit and Wolf introduced a well-conditioned and consistent linear estimator of the sample covariance in [13] and a considerably more complicated nonlinear one in [14]. Here we employ Ledoit and Wolf's simple linear estimator.

$$\tilde{\mathbf{\Sigma}}_x = \lambda m_x\mathbf{I} + (1-\lambda)\mathbf{S}_x,$$

where $\mathbf{S}_x$ is the sample covariance, m_x is the average eigenvalue of $\mathbf{S}_x$, and $\lambda \in [0, 1]$ is a convex mixing coefficient that shrinks the sample covariance towards $m_x\mathbf{I}$ as λ approaches 1. Ledoit and Wolf derive a value for λ to ensure that $\tilde{\mathbf{\Sigma}}_x$ is a consistent estimator of the true covariance $\mathbf{\Sigma}_x$.

We apply both PMD and our extension of it on the simulated data described above. As a proof of concept - to see if we could recover the generative sparse model - we hand picked the regularization parameters λ_a and λ_b to see if there was a pair of values for which we could recover the true set of SNPs. In particular, we are interested in how many relevant SNPs were missed when sufficient regularization was applied to drop all irrelevant SNPs from the model. In practice, we would choose the regularization parameters with either a measure of complexity such as the BIC or by a data driven method such as cross-validation.

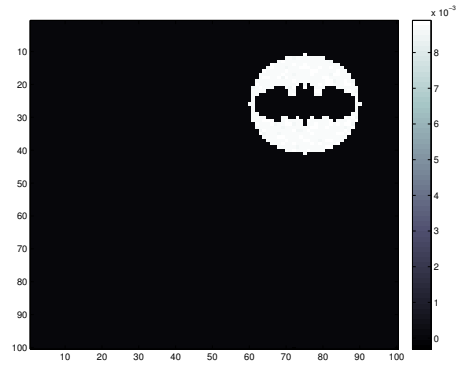

Fig. 3. With non-trivial covariance estimate: The unfolded vector $\mathbf{b}$ that summarizes $\mathbf{Y}$.

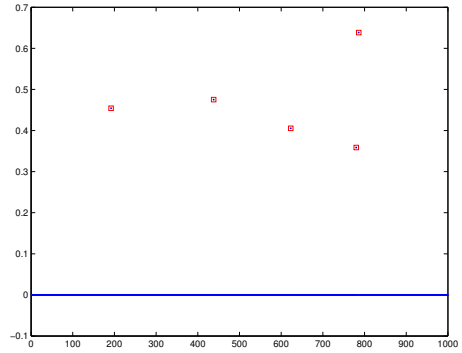

Fig. 4. With non-trivial covariance estimate: The estimated sparse vector $\mathbf{a}$ that summarizes $\mathbf{X}$. The SNP loci annotated in red denote the loci used to generate the data.

3. RESULTS

Figures 3 and 4 show the estimated canonical correlation vectors $\mathbf{b}$ and $\mathbf{a}$ unfolded when the non-trivial covariance estimate is used. We see that there is a regularization parameter that recovers the correct support. Figure 5 shows estimated $\mathbf{a}$ obtained via PMD using hand picked regularization parameters. PMD selected the same set of voxels and for space considerations, the results are not shown. Nonetheless, interestingly, the selected SNPs are different. Choosing a smaller λ_a will indeed include the missed SNP, but the cost is that false positives will also be included.

4. DISCUSSION

In this paper we build on previous penalized multivariate methods for finding sparse structure in pairs of related data sets by showing how to incorporate correlation information. Our simulation example shows that our method is capable of recovering true latent sparse structure and that the solutions obtained when accounting for correlation structure can differ from multivariate approaches that assume identity co-

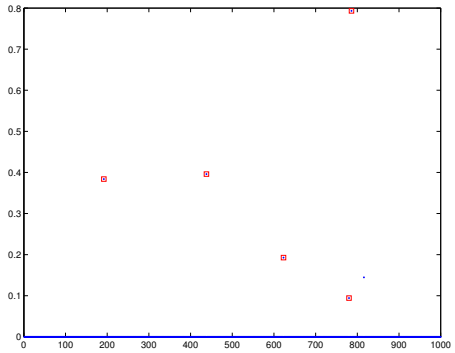

Fig. 5. PMD: The estimated sparse vector $\mathbf{a}$ that summarizes $\mathbf{X}$. The SNP loci annotated in red denote the loci used to generate the data.

variances. Using non-trivial covariance estimates, however, makes the optimization problem harder. To that end we are working on developing more efficient algorithms that can work with non-trivial covariance matrices. Additionally, we are currently investigating our methods on real data.

5. REFERENCES

[1] Paul M Thompson, Nicholas G Martin, and Margaret J Wright, "Imaging genomics," *Current Opinion in Neurology*, vol. 23, no. 4, 2010.

[2] Alexander H. Joyner, Cooper Roddey J., Cinnamon S. Bloss, Trygve E. Bakken, Lars M. Rimol, Ingrid Melle, Ingrid Agartz, Srdjan Djurovic, Eric J. Topol, Nicholas J. Schork, Ole A. Andreassen, and Anders M. Dale, "A common mecp2 haplotype associates with reduced cortical surface area in humans in two independent populations," *Proceedings of the National Academy of Sciences*, vol. 106, no. 36, pp. 15483–15488, 2009.

[3] Steven G. Potkin, Jessica A. Turner, Guia Guffanti, Anita Lakatos, James H. Fallon, Dana D. Nguyen, Daniel Mathalon, Judith Ford, John Lauriello, Fabio Macciardi, and FBIRN, "A genome-wide association study of schizophrenia using brain activation as a quantitative phenotype," *Schizophrenia Bulletin*, vol. 35, no. 1, pp. 96–108, 2009.

[4] Nicola Filippini, Anil Rao, Sally Wetten, Rachel A. Gibson, Michael Borrie, Danilo Guzman, Andrew Kertesz, Inge Loy-English, Julie Williams, Thomas Nichols, Brandon Whitcher, and Paul M. Matthews, "Anatomically-distinct genetic associations of apoe 4 allele load with regional cortical atrophy in alzheimer's disease," *NeuroImage*, vol. 44, no. 3, pp. 724 – 728, 2009.

[5] Jason L. Stein, Xue Hua, Suh Lee, April J. Ho, Alex D. Leow, Arthur W. Toga, Andrew J. Saykin, Li Shen, Tatiana Foroud, Nathan Pankratz, Matthew J. Huentelman, David W. Craig, Jill D. Gerber, April N. Allen, Jason J. Corneveaux, Bryan M. DeChairo, Steven G. Potkin, Michael W. Weiner, and Paul M. Thompson, "Voxelwise genome-wide association study (vgwas)," *NeuroImage*, vol. 53, no. 3, pp. 1160 – 1174, 2010.

[6] Harold Hotelling, "Relations between two sets of variants," *Biometrika*, vol. 28, pp. 321–377, 1936.

[7] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[8] Daniela M. Witten, Robert Tibshirani, and Trevor Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.

[9] Maria Vounou, Thomas E. Nichols, and Giovanni Montana, "Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach," *NeuroImage*, vol. 53, no. 3, pp. 1147–1159, 2010.

[10] Maria Vounou, Eva Janousova, Robin Wolz, Jason L. Stein, Paul M. Thompson, Daniel Rueckert, and Giovanni Montana, "Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in alzheimer's disease," *NeuroImage*, vol. 60, no. 1, pp. 700 – 716, 2012.

[11] H. D. Vinod, "Canonical ridge and econometrics of joint production," *Journal of Econometrics*, vol. 4, no. 2, pp. 147–166, May 1976.

[12] Ignacio González, Sébastien Déjean, Pascal G. P. Martin, and Alain Baccini, "Cca: An r package to extend canonical correlation analysis," *Journal of Statistical Software*, vol. 23, no. 12, pp. 1–14, 1 2008.

[13] Olivier Ledoit and Michael Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365 – 411, 2004.

[14] Olivier Ledoit and Michael Wolf, "Nonlinear shrinkage estimation of large-dimensional covariance matrices," *Annals of Statistics*, vol. 40, no. 2, pp. ,1024–1060, 2012.