An integration of partial least squares and sparsified multiple correspondence analysis applied to genetic data

Derek Beaton[a*], Anne Bernard[b], Hervé Abdi[a], and Gilbert Saporta[c]

[a]The University of Texas at Dallas, USA
[b] University of Queensland, Australia
[c] CNAM Paris, France

Keywords: Genetics, behavioral data, sparsification, multiple correspondence analysis, group sparsification

Abstract

Genetic data are now common in many domains. Typically, these genetic studies try to associate genetics with single phenotypes, behaviors, or diagnostic criteria. However, many of these studies include multiple behavioral variables and very large genetic data sets. The analysis of these data sets faces two particular challenges: 1) How to integrate many behavioral and genetic variables when 2) only a small number of variables are interpretable. To address these issues, we propose the integration of partial least squares and a sparsified approach to multiple correspondence analysis.

1 Introduction

The singular value decomposition (SVD)—and its weighted least-squares extension, the generalized SVD (GSVD)—are the core of many multivariate techniques such as principal components analysis (PCA), partial least squares (PLS), and multiple correspondence analysis (MCA). All of these techniques produce components (a.k.a. latent variables) which are obtained as linear combinations of the original variables. However, when the sample size is small and the data large (*i.e.,* $n << p$) many variables will have non-zero loadings, a pattern that makes it difficult to interpret the results. A number of regularization and sparsification techniques have been proposed in order to produce only few non-zero loadings. Some of these methods include SCoTLASS [1], SPCA [2], SPCA-rSVD[3], and rPCA[4].

However, with large genetics data there are still some problems not easily addressed by sparsification alone. In fact, using behavioral data in conjunction with large genetics data (such as genome-wide data) may increase power to detect genetic effects [5]. Therefore, it would be advantageous to use a PLS method designed for behavioral and genetic data. One such method is PLSCA [6]. But, PLSCA would still produce many non-zero loadings. Because most genetic data (such as SNPs) are categorical they are naturally structured by blocks (see Table 3). A recent extension of SPCA-rSVD[3] has been adapted as group-sparse PCA and extended to MCA, called sparse MCA (SMCA) [7].

In this paper we present a solution to the analysis of genetic data that integrates PLSCA with SMCA. This approach can associate SNPs—typically large and noisy data—to various behavioral markers—typically well-defined instruments—while producing as few non-zero loadings as possible.

Table 3: Example of nominal (left) and disjunctive (right) coding of illustrative SNPs (SNP 1 and 2)

	SNP 1	SNP 2
*Subj.*1	AG	CA
...	...	...
Subj.I	GG	AA

(a) Nominal

	SNP 1			SNP 2		
	AG	AA	GG	CA	AA	CC
*Subj.*1	1	0	0	1	0	0
...	...	...	...	...	...	...
Subj.I	0	0	1	0	1	0

(b) Disjunctive

2 Asymmetric sparsification of genetic data

Call $\mathbf{Y}$ a disjunctive data matrix that represents behavioral data (e.g., a survey). Call $\mathbf{X}$ a disjunctive data matrix that represents genetic data (e.g., SNPs; see, e.g., Table 3). SNPs can be viewed as multiblock data, where each SNP represents a block of (usually) 3 columns. We would approach the analysis of these data sets with PLSCA [6] as such: $\mathbf{R} = \mathbf{Y}^{\mathsf{T}}\mathbf{X}$ where $\mathbf{R}$ is a contingency table (behavioral data $\times$ genetic data). Correspondence analysis (CA) is a natural choice for analyzing very large contingency tables. Therefore, in PLSCA, we would preprocess $\mathbf{R}$ as we normally would with CA, where the row and column weights are proportional to the row and column sums and stored in diagonal matrices ($\mathbf{W_Y}$ and $\mathbf{W_X}$, respectively). We then decompose $\mathbf{R}$ with the GSVD where $\mathbf{R} = \mathbf{P\Delta Q}^{\mathsf{T}}$, such that $\mathbf{PMP} = \mathbf{QWQ} = \mathbf{I}$, where the latent variables are: $\mathbf{L_X} = \mathbf{XW_XP} \times r_{++}^{\frac{1}{2}}$ and $\mathbf{L_Y} = \mathbf{YW_YQ} \times r_{++}^{\frac{1}{2}}$ where r_{++} is the sum of the table $\mathbf{R}$. The latent variables have maximal covariance (due to the properties of the GSVD) as: $\mathbf{L_X}^{\mathsf{T}}\mathbf{L_Y} = \mathbf{\Delta}$

However, there are typically many SNPs in data such as these, thus implying that $\mathbf{X}$ has the following two properties: 1) a natural block structure and 2) is much larger than $\mathbf{Y}$ and thus could produce components with many non-zero values. If $\mathbf{X}$ has a natural block structure such that $\mathbf{X} = [\mathbf{X}_1, ..., \mathbf{X}_b, ..., \mathbf{X}_B]$ where $\mathbf{R} = \mathbf{Y}^{\mathsf{T}}[\mathbf{X}_1, ..., \mathbf{X}_b, ..., \mathbf{X}_B]$ and given the block structure of $\mathbf{R}$, we have: $\mathbf{R} = \mathbf{P\Delta Q}^{\mathsf{T}} = \mathbf{P\Delta}[\mathbf{Q}_1, ..., \mathbf{Q}_b, ..., \mathbf{Q}_B]^{\mathsf{T}}$. The blocks of $\mathbf{Q}$ represent the group coding of SNPs, and hence, we would want to (group) sparsify the SNPs (genetic) data. Sparsification can be achieved with a extension of group sparse PCA to MCA: sparse multiple correspondence analysis [7].

We will present the details of the method illustrated with a small and a realistic example.

References

[1] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin, "A modified principal component technique based on the lasso," *Journal of Computational and Graphical Statistics* **12**(3), pp. 531–547, 2003.

[2] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics* **15**(2), pp. 265–286, 2006.

[3] H. Shen and J. Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *Journal of multivariate analysis* **99**(6), pp. 1015–1034, 2008.

[4] M. Verbanck, J. Josse, and F. Husson, "Regularised PCA to denoise and visualise data," *Statistics and Computing* , pp. 1–16, 2013.

[5] E. D. Schifano, L. Li, D. C. Christiani, and X. Lin, "Genome-wide association analysis for multiple continuous secondary phenotypes," *American journal of human genetics* **92**(5), pp. 744–759, 2013.

[6] D. Beaton, F. M. Filbey, and H. Abdi, "Integrating partial least squares correlation and correspondence analysis for nominal data," in *New Perspectives on Partial Least Squares and Related Methods*, H. Abdi, W. Chin, V. Esposito Vinzi, G. Russolillo, and L. Trinchera, eds., *Proc. in Mathematics and Statistics* **56**, pp. 81–94, 2013.

[7] A. Bernard, C. Guinot, A. Tenehaus, H. Abdi, and G. Saporta, "Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis," *NA* , In Prep.

*Corresponding author. E-mail: derekbeaton@utdallas.edu.