Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications Note

OXFORD

Genetics and population analysis

Fast sparse canonical correlation analysis with flashpca

Gad Abraham [1,2]* and Michael Inouye [1,2]

[1] Centre for Systems Genomics, School of BioSciences, University of Melbourne, Parkville 3010, VIC, Australia and
[2] Department of Pathology, Faculty of Medicine, Dentistry, and Health Sciences, University of Melbourne, Parkville 3010, VIC, Australia.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Sparse canonical correlation analysis (SCCA) is a useful approach for correlating one set of measurements, such as single nucleotide polymorphisms (SNPs), with another set of measurements, such as gene expression levels. We present a fast implementation of SCCA, enabling rapid analysis of hundreds of thousands of SNPs together with thousands of phenotypes. Our approach is implemented both as an R package `flashpcaR` and within the standalone command-line tool `flashpca`.

Availability and implementation: https://github.com/gabraham/flashpca
Contact: gad.abraham@unimelb.edu.au
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Canonical correlation analysis (CCA) is a well-known statistical method for multivariate analysis of two datasets (Hotelling, 1936), such as genotype data (single nucleotide polymorphisms) and multivariate phenotypes such as gene expression levels. Approaches that consider one SNP at a time together with multiple phenotypes have been shown to increase power to detect QTLs over the simpler but often-used single-SNP/single-phenotype approach (Ferreira *et al.*, 2009), particularly when analysing correlated phenotypes that are modulated by the same genetic variation.

Analysis of multiple SNPs at a time is an attractive extension of approach, however, standard CCA is not well-defined in this case, as typically the number of samples is substantially lower than the number of SNPs ($n \ll p$). One solution is Sparse CCA (SCCA) (Witten *et al.*, 2009b,a; Parkhomenko *et al.*, 2009), an L_1-penalised variant of CCA where the number of variables that effectively contribute to the canonical correlation can be tuned, making the problem well-defined. Owing to the induced sparsity, SCCA can be useful for identifying a small subset of SNPs and a small subset of the phenotypes exhibiting strong correlations. However, the rapidly increasing size and coverage of genotyping arrays (exacerbated by genotype imputation), together with the availability of large phenotypic datasets (transcriptomic, metabolomic, and others), makes it challenging to perform analyses such as SCCA using existing tools.

We have developed an efficient implementation of SCCA that is capable of analysing genome-wide SNP datasets (1M SNPs or more), together with thousands of phenotypes (such as gene expression levels), as part of the tool `flashpca` (Abraham *et al.*, 2014). The main code is implemented in C++ using the Eigen 3 numerical library (Guennebaud *et al.*, 2010), and an R implementation (package `flashpcaR`) is available via RcppEigen (Bates *et al.*, 2013).

Here, we compare the SCCA implementation in `flashpcaR` and `flashpca` with a widely-used implementation (`PMA`, by Witten *et al.* (2013)), and demonstrate the substantial improvements in speed of our tool, while achieving comparable or better cross-validated predictive power.

2 Methods

In standard CCA, we assume that we have two matrices $\mathbf{X}$ ($n \times p$) and $\mathbf{Y}$ ($n \times m$), measured for the same n samples. We further assume that both $\mathbf{X}$ and $\mathbf{Y}$ have been column-wise standardised (zero mean, unit variance). For a single pair of canonical variables a and b, CCA involves solving the problem

$$\underset{a,b}{\arg\max} \frac{a^T \mathbf{\Sigma}_{XY} b}{\sqrt{a^T \mathbf{\Sigma}_{XX} a \; b^T \mathbf{\Sigma}_{YY} b}}, \tag{1}$$

where $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{YY}$ are the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and $\mathbf{\Sigma}_{XY}$ is the covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$. The solution is given by the singular value decomposition of $\mathbf{\Sigma}_{XX}^{-1/2} \mathbf{\Sigma}_{XX} \mathbf{\Sigma}_{XY}^{-1/2}$, with

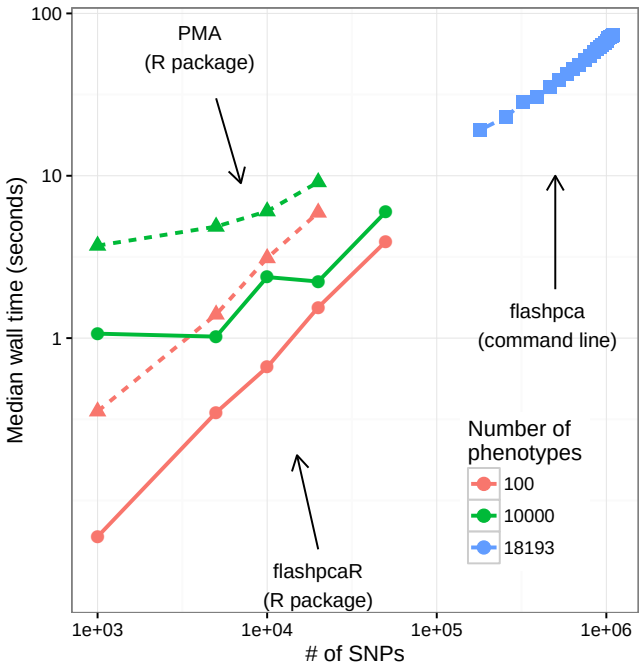

Fig. 1. Timing (median of 30 runs) of SCCA implemented in (i) the `flashpcaR` (R package) and (ii) `flashpca` (stand-alone command-line tool), compared with SCCA from `PMA`, using subsets of the HapMap3 dataset with gene expression levels as phenotypes.

$a = \mathbf{\Sigma}_{XX}^{-1/2} u_1$ and $b = \mathbf{\Sigma}_{YY}^{-1/2} v_1$, where u_1 and v_1 are the first left and right singular vectors.

SCCA is typically run on high-dimensional data, where a useful assumption is that the variables within $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated, i.e., $\mathbf{\Sigma}_{XX} = \mathbf{\Sigma}_{YY} = \mathbf{I}$ (Parkhomenko *et al.*, 2009), hence, $a = u$ and $b = v$. Thus, SCCA involves solving a simplified form of CCA,

$$\underset{u,v}{\arg\max} \quad u^T \mathbf{\Sigma}_{XY} v$$
$$\text{s.t. } ||u||_2^2 = 1, ||v||_2^2 = 1, ||u||_1 \leq s_u, ||v||_1 \leq s_v, \qquad (2)$$

where u and v are the left and right canonical vectors, and s_u and s_v are constraints on the L_1 norms of canonical vectors.

The problem can be converted into the penalised (Lagrangian) form, making it solvable using iterative soft-thresholding (Parkhomenko *et al.*, 2009), which we employ here. Unlike standard CCA, SCCA is well-defined even when $n < \min\{p, m\}$, and induces sparse canonical vectors, depending on the choice of L_1 penalties: higher penalties lead to higher sparsity. The optimal set of penalties can be found via cross-validation: the data (both $\mathbf{X}$ and $\mathbf{Y}$) are split into training and test sets, SCCA is run on the training set $(\mathbf{X}_{train}, \mathbf{Y}_{train})$ using a 2D grid of penalties, and the pair of penalties that produce the highest correlations in the test set, $\text{Cor}(\mathbf{X}_{test} u, \mathbf{Y}_{test} v)$, are selected. Optionally, a new model may be fitted to the entire data using these penalties.

3 Results

To demonstrate our tool we utilised the HapMap3 phase III genotypes (International HapMap 3 Consortium, 2010), together with gene expression data of 602 individuals (Stranger *et al.*, 2012), with 1.4M SNPs and 47,000 gene expression probes (out of which 21,800 probes were analysed by Stranger *et al.* (2012) and used in our analysis). After quality control (see Supplementary Material), our data consisted of 601 individuals, 1,125,747 SNPs, and 18,193 gene expression probes.

We first confirmed that `flashpcaR::scca` produced models competitive with `PMA::CCA`, by comparing the best Pearson correlations produced by each model in cross-validation. Using chromosome 1 from the HapMap3 data (89,603 SNPs), together with the 18,193 gene expression levels, we performed 5-fold cross-validation with a 2D grid of penalties (see the Supplementary Material for details). The maximum of the average test-fold Pearson correlations was identical for both `flashpcaR::scca` and `PMA::CCA` ($\rho = 0.044$).

Next, we compared the speed of three tools: (i) `PMA::CCA` (R package), (ii) `flashpcaR::scca` (R package), and (iii) `flashpca` (stand-alone command-line tool). Figure 1 shows a timing comparison between `flashpcaR::scca` and `PMA::CCA` (using pre-computed SVD of $\mathbf{X}^T\mathbf{Y}$ via `flashpcaR::flashpca`). `flashpcaR::scca` was $\sim$9$\times$ faster than `PMA::CCA`, with a medium-sized analysis (20,000 SNPs and 10,000 gene expression levels) completing in 1s for the former but 9s for the latter. For performing a 20×20-penalty grid search in 10-fold cross-validation (assuming a single core), this would translate to $\sim$1h for `flashpcaR` versus $\sim$9h for `PMA`.

Whereas both `PMA::CCA` and `flashpcaR::flashpca` are bound by the memory limitations of R, the command-line tool `flashpca` allows much larger analyses. We ran analyses of increasing size: chromosomes 1-2, 1-3, . . . , 1-22, up to all 1,125,747 SNPs. The stand-alone `flashpca` was able to complete an analysis of 601 individuals, 1,125,747 SNPs and 18,193 gene expression levels in $\sim$30sec (median of 30 runs), using $\sim$9GiB of RAM.

4 Conclusion

`flashpca` provides a fast implementation of sparse canonial correlation analysis, making it possible to rapidly analyse high dimensional datasets. For datasets too large to fit in R, the command-line tool is available as well, enabling fast analysis of >1M SNPs and thousands of phenotypes at a time. In addition to canonical correlation analysis of multiple quantitative phenotypes, fast sparse partial least squares (sparse PLS) can be performed by using a single phenotype at a time.

Funding

This work has been supported by the NHMRC grant no. 1062227. GA was supported by an NHMRC Early Career Fellowship (no. 1090462). MI was supported by a Career Development Fellowship co-funded by the NHMRC and Heart Foundation (no. 1061435).

References

Abraham, G. *et al.* (2014). Fast principal component analysis of large-scale genome-wide data. *PLoS ONE*, **9**(4), e93766.

Bates, D. *et al.* (2013). Fast and elegant numerical linear algebra using the RcppEigen package. *Journal of Statistical Software*, **52**(5), 1–24.

Ferreira, M. A. R. *et al.* (2009). A multivariate test of association. *Bioinformatics*, **25**, 132–133.

Guennebaud, G. *et al.* (2010). Eigen v3. http://eigen.tuxfamily.org.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, **28**, 321–377.

International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.

Parkhomenko, E. *et al.* (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical applications in genetics and molecular biology*, **8**(1), Article 1.

Stranger, B. E. *et al.* (2012). Patterns of Cis Regulatory Variation in Diverse Human Populations. *PLoS Genetics*, **8**(4), e1002639.

Witten, D. *et al.* (2009a). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–34.

Witten, D. *et al.* (2009b). Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. *Statistical applications in genetics and molecular biology*, **8**(1), Article 29.

Witten, D. *et al.* (2013). *PMA: Penalized Multivariate Analysis*. R package version 1.0.9.