Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach

Maria Vounou [a], Thomas E. Nichols [b], Giovanni Montana [a,*]
and the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Statistics Section, Department of Mathematics, Imperial College London, UK
[b] Department of Statistics & Warwick Manufacturing Group, University of Warwick, UK

ARTICLE INFO

Article history:
Received 31 March 2010
Revised 24 June 2010
Accepted 1 July 2010
Available online xxxx

ABSTRACT

There is growing interest in performing genome-wide searches for associations between genetic variants and brain imaging phenotypes. While much work has focused on single scalar valued summaries of brain phenotype, accounting for the richness of imaging data requires a brain-wide, genome-wide search. In particular, the standard approach based on mass-univariate linear modelling (MULM) does not account for the structured patterns of correlations present in each domain. In this work, we propose sparse reduced rank regression (sRRR), a strategy for multivariate modelling of high-dimensional imaging responses (measurements taken over regions of interest or individual voxels) and genetic covariates (single nucleotide polymorphisms or copy number variations), which enforces sparsity in the regression coefficients. Such sparsity constraints ensure that the model performs simultaneous genotype and phenotype selection. Using simulation procedures that accurately reflect realistic human genetic variation and imaging correlations, we present detailed evaluations of the sRRR method in comparison with the more traditional MULM approach. In all settings considered, sRRR has better power to detect deleterious genetic variants compared to MULM. Important issues concerning model selection and connections to existing latent variable models are also discussed. This work shows that sRRR offers a promising alternative for detecting brain-wide, genome-wide associations.

1. Introduction

Recent attention in imaging neuroscience has been focused on the genome-wide search for genetic variants that explain the variability observed in both brain structure and function. In this sense, the field of *imaging genetics* is catching up with the dramatic increase in the number of genome-wide association (GWA) studies that have been reported across many different disease areas and that have been fuelled by recent technological improvements in genotyping and reductions in cost.

The fundamental assumption that underlies the GWA approach is that extensive common variation in the human genome, as measured by single nucleotide polymorphisms (SNPs) or copy number variations (CNVs) for example, contribute to the risk of most common disorders. Over the last few years, substantial international resources have been directed in an effort to better characterise human genetic variation, for instance through the HapMap[2] and the Genome 1000 projects.[3] Nonrandom association or *linkage disequilibrium* (LD) between alleles at nearby loci, means that not all loci in a chromosomal region need be genotyped for the majority of common variation to be captured, so that the spacing between markers should only be dense enough to capture the variation at those loci that have not been genotyped.

The latest genotyping platforms enable the measurement of around 1.8 million genetic markers, including SNPs and CNVs, enabling a search for statistically significant associations between one or more markers and the phenotype. Depending on the study design, the phenotype is usually encoded as a dichotomous variable (e.g., as a case or control) or as a quantitative trait, either univariate or multivariate. The belief is that variants yielding an increase in disease risk will be more easily found by means of such population-based association studies, as compared with alternative approaches such as family-based linkage analysis studies. For binary phenotypes, recent studies have identified significantly associated SNPs that are in LD with predisposing variants that increase the disease risks by between

[*] Corresponding author.
E-mail address: g.montana@imperial.ac.uk (G. Montana).

[1] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete list of ADNI investigators is available at http://www.loni.ucla.edu/ADNI/Collaboration/ADNI_Manuscript_Citations.pdf.

[2] http://snp.cshl.org.

[3] http://www.1000genomes.org.

1053-8119/$ – see front matter © 2010 Published by Elsevier Inc.
doi:10.1016/j.neuroimage.2010.07.002

10% and 30% over noncarriers (Donnelly, 2008). A concern is that many more common variants may not have been detected in GWA studies because they contribute to raising the risk by much smaller proportions.

A number of population-based association studies with neuroimaging phenotypes have appeared in the literature over the last few years. Depending on both the dimensionality of the phenotype being investigated and the size of genomic regions being searched for association, we can attempt a broad classification of the existing imaging genetic studies into four main categories. Some studies can be classified as belonging to the *candidate phenotype–candidate gene association* (CP–CGA) category, meaning that a specific gene or chromosomal region is tested for association with a typically low-dimensional phenotype. The assumption is that the particular quantitative phenotypes being measured is able to capture changes in the brain induced by the disease or other biological condition being studied. An example of this approach is described by Joyner et al. (2009), who examined the potential association between 4 summary brain structure measures used as surrogate of brain size and 11 SNPs located in and around the MECP2 gene. They studied two different populations—a homogeneous population consisting of healthy controls and patients with psychotic disorders, and a heterogenous population of healthy controls and patients with mild cognitive impairment. Other studies belong to the *candidate phenotype–genome-wide association* (CP–GWA) category where, again, the phenotype has been appropriately identified but the search for genetic variants has a much wider scope. An example is given by Potkin et al. (2008), who use a brain imaging activation signal in the dorsolateral prefrontal cortex as the quantitative trait reflecting schizophrenia dysfunction, and present a genome-wide study based on subjects with chronic schizophrenia and controls matched for gender and sex. Other studies have taken the opposite approach and fall into the *brain-wide, candidate-gene association* (BW-CGA) class. In this case, the search for genetic variants is confined to specific chromosomes or regions of interest but is extended to the entire brain by means of very high-dimensional phenotypes, typically based on voxel-based morphometry techniques. Filippini et al. (2009) described one such study, in which a whole-brain search for associations between the ApoE $\varepsilon4$ allele load and gray matter volume in the entire brain is carried out by testing for both additive and genotypic models in a large mild AD population.

We predict that soon GWA studies in neuroimaging genetics will embrace the *brain-wide, genome-wide association* (BW-GWA) paradigm, where both the entire genome and entire brain are searched for nonrandom associations and other interesting dependence patterns. BW-GWA studies necessarily rely on very high-dimensional phenotypes. The assumption is that only a handful of quantitative traits (e.g., voxels or voxel clusters) may be found in a statistically meaningful association with a handful of genetic markers. The approach requires a statistical framework for the simultaneous identification of *localised* genomic regions and *localised* brain regions that are found to be in nonrandom association. A very recent example is the study carried out by Stein et al. (2010). Here a voxel-wise search for variants that influence brain structure was performed, using approximately 44,8000 single nucleotide polymorphisms and around 31,000 voxels across the entire brain . In this article, we focus on both computational and statistical issues arising in a BW-GWA study. Consider the case with p genetic markers and q quantitative phenotypes, with both p and q being much smaller than the available sample size n. A simple modelling approach consists of fitting all possible $(p \times q)$ univariate linear regression models, all independently of each other, and ranking genotype–phenotype pairs by p-value. This approach, often referred to as *mass-univariate linear modelling* (MULM), is appealing because of its simplicity and because univariate regression models can be easily fitted even when only small sample sizes are available. However, despite its advantages, it presents at least three major shortcomings.

The first limitation is related to the need, typical of a mass-univariate GWA study, to determine an experiment-wide significance level that accounts for the multiple testing problem. Whether a family-wise error or false discovery rate approach is used, the complex dependence structure among both genetic markers and among phenotypes must be accounted for. For example, Stein et al. (2010) collapse inferences over the p SNPs at each voxel by taking the minimum P-value, and then corrects for the effective dimensionality accounting for LD. Other approaches rely on computationally intensive permutation procedures.

A second important limitation of MULM is that it does not exploit the possible spatial structure of phenotype–genotype associations. If a genetic marker explains phenotype variance at one brain location, we expect it will likely affect other neighbouring locations as well. Hence, we would expect that an association mapping approach that is able to ‘borrow strength’ from correlated phenotypes can potentially yield higher statistical power (Ferreira and Purcell, 2009).

Lastly, MULM does not account for the possibility that multiple markers, possibly located on different genes, may jointly contribute to a particular phenotypic effect. In this instance, a multivariate approach that combines genetic information from multiple markers simultaneously into the analysis is also expected to provide greater power (Kwee et al., 2008).

In an attempt to address these shortcomings, we derive a new statistical methodology for multilocus mapping in BW-GWA studies. Our novel approach is based on regularised (or penalised) regression techniques, a class of regression models offering a natural way of searching simultaneously for multiple markers that are highly predictive of phenotype. Penalised regression has recently been described as promising alternative to more traditional SNP ranking and hypothesis testing procedures (Cantor et al., 2010). Penalised regression methods are particularly suitable where $p>>n$ since they perform ‘model selection’, highlighting subsets of predictors that demonstrate greatest effect on the response. Penalised regression works by estimating the regression coefficients in the linear model, subject to constraints. Examples include ridge regression and Lasso regression (Tibshirani, 1996). Specifically, the Lasso estimator solves the ordinary least squares problem when a penalisation on the L1 norm of the coefficients is added to the mean square error objective function. Depending on the degree of penalisation, Lasso regression drives some coefficients exactly to zero, excluding them from the model, and thus performing variable selection. In the context of GWA studies, sparse generalised linear models, and specifically logistic regression, have been used to select genetic markers that are highly predictive of the disease status (Cantor et al., 2010; Croiseau and Cordell, 2009; Hoggart et al., 2008; Wu et al., 2009).

In this article, we extend this approach to accommodate high-dimensional quantitative responses, such that both *covariate selection* and *response selection* can be performed simultaneously. The proposed approach, *sparse reduced–rank regression*, performs both genotype and phenotype selection required by BW-GWA studies, and is computationally less expensive than the mass-univariate approach.

To compare the power of our method to that of conventional MULM, we introduce a detailed simulation framework that associates a small number of markers with gray matter volume. We use a realistic simulation of both genomic and phenotypic variation. Further realism is introduced by subsequently removing true causative markers from the study, so that genotype–phenotype associations can be detected only through markers that are in LD with these excluded markers. To the best of our knowledge, our extensive simulation results provide a first characterisation of the statistical power of BW-GWA *imaging genetics* studies, for both univariate and multivariate approaches.

2. Materials and methods

2.1. Data simulation procedure

We have developed a realistic simulation framework for assessing the performance of any statistical approach for population-based

Please cite this article as: Vounou, M., et al., Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.07.002

association mapping with neuroimaging phenotypes. Our simulation procedure initially generates genomes that make up a large human population. We used the FREGENE genome simulator to generate a large population of human genomes. The simulation process evolves the population forwards in time, over several nonoverlapping generations, by keeping track of complete ancestral information. The simulations are set up so as to reproduce the effects of salient evolutionary forces, such as mutation, recombination, and selection, with parameters chosen to mimic the evolutionary processes inferred from real human populations. At the end of the simulation, each genome in the population is represented by a high-dimensional vector of biallelic genetic markers, that is then paired up with multivariate neuroimaging vector derived from real MRI data using VBM. Finally, a precise statistical association linking a handful of genotypes and a handful of phenotypes is induced in the population by carefully modifying the quantitative phenotypes according to a genetic model.

From this large target population, repeated random samples of any size can be extracted. For each sample, the true underlying genotype–phenotype dependence is known, and the performance of any statistical method for detecting genetic associations can be easily assessed. The use of data simulated under a predetermined genetic model enables us to study the performance of competing statistical models in an unbiased fashion by means of performance measures such as ROC (receiver operating characteristic) curves, which would otherwise be impossible to evaluate in real studies. Our approach also provides a framework for characterising the statistical power required to detect true, nonrandom associations. A detailed description of our simulation and calibration procedures is provided below.

2.1.1. Genotype simulation

The simulation of a large human population was carried out using the simulation software FREGENE (FoRward Evolution of GENomic rEgions) (Hoggart et al., 2007). The software implements a forward-in-time simulation procedure in which each individual's genome consists of a single linear chromosome having minor allele counts. The population evolves over nonoverlapping generations according to a Wright–Fisher model, with specific control over the population genetic parameters including selection coefficients, recombination, migration rates, population size, and structure. Using FREGENE, we initially generated a panmictic human population that mimics the evolution of $N=10{,}000$ diploid individuals along 200,000 generations. We used a per-site mutation rate of 2.3×10^{-8}, a per-site crossover rate of 1.1×10^{-8}, and a per-site gene conversion rate of 4.5×10^{-9}, with 80% of recombination events occurring in hotspots, with a 2-kb hotspot length. Selection was also introduced, with the proportion of sites under selection set to 5×10^{-4}. Each simulated sequence was 20 Mb long. Since each marker is biallelic, we will denote the two alleles as *A* and *a*, with genotypes text *AA*, *Aa*, or *aa*. For each SNP, the minor allele frequency (MAF) is then $f_{aa}+f_{Aa}/2$ where f_{aa} and f_{Aa} are the population frequencies of genotypes *aa* and *Aa*. The genotype for individual *i* at locus *s* is denoted by x_{is} $(i=1,\ldots,N, s=1,\ldots,p)$ and represents the count of minor allele recorded at that locus (homozygote of minor allele is 2, heterozygote is 1, and homozygote of major allele is 0). SNPs having a MAF smaller than 0.05 were initially removed, leaving a total of $p=37748$ markers. Of these, $k=10$ markers having MAF = 0.2 were preselected to act as causative SNPs—these were randomly chosen only once and held fixed in all subsequent simulations and analyses. The causative SNPs are only used to introduce genetic effects on the phenotypes (see below for details) and are removed from each data set before any statistical analysis.

2.1.2. Data, MRI analysis, and phenotype simulation

Brain phenotype simulations were generated using MRI data obtained from the publicly available Alzheimer Disease Neuroimaging Initiative (ADNI) database.[4] The primary goal of ADNI is to test whether serial imaging and nonimaging measures can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer disease (AD). Data are collected at a range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the United States and Canada. Complete background and methodological detail of the ADNI data can be found on the project Web site.[5] For our study, we only used baseline T1 MRI scans from 189 subjects with MCI. The ADNI T1 MRI scans have an initial resolution of $0.9375\times0.9375\times1.2$ mm^3 (3D MP-RAGE sequence, TR = 2400 ms, TE = 1000 ms, FA = 8°) and were preprocessed with the SPM5 'optimised' VBM procedure (Good et al., 2001), using a unified segmentation and warping method, followed by modulation of gray matter (GM) segmented images by the Jacobian of the warping. This produces GM images in standard space that still retain units of GM volume of the individual. The resulting images, $2.0\times2.0\times2.0$-mm^3 in resolution in the MNI space, were used with no applied smoothing.

From each image, we extracted the mean modulated GM value from $q=111$ anatomical ROIs defined by the GSK CIC Atlas (Tziortzi et al., 2010). The GSK CIC Atlas is based on the Harvard-Oxford atlas[6] but offers a 6-level hierarchy, from a coarse 3-region (gray matter, cerebral white matter, and CSF) version to a fine 111-region version (illustrated in Fig. 1). After regressing out the effect of gender and age, we estimated the ROI means, all collected in a vector $\mu=(\mu_1,\mu_2,\ldots,\mu_q)$, and their covariance matrix Σ. For each individual *i* in the simulated population, we generated imaging phenotypes by simulating a vector $\mathbf{y}_i=(y_{i1},y_{i2},\ldots,y_{iq})$ drawn from the multivariate normal distribution with parameters (μ,Σ). The values in $\mathbf{y}_i$ can be interpreted as baseline GM measurements, unlinked to genotypes, before the introduction of genetic effects.

2.1.3. Genetic effects

We induced genetic effects in $l=6$ ROIs using an additive genetic model involving the $k=10$ causative SNPs. To simplify notation, we let the first *k* genotypes correspond to the causal SNPs, and the first *l* phenotypes correspond to the affected ROIs. Recalling that y_j is the simulated baseline GM value for ROI *j*, the target phenotypes have their GM intensity reduced as per

$$y_j^* = y_j - w_j$$

where

$$w_j = \delta_j \sum_{s=1}^{k} \zeta_{js} x_s \quad \text{subject to} \quad \sum_{s=1}^{k} \zeta_{js} = 1$$

for $j=1,\ldots,l$. Each w_j term represents the reduction due to the additive genetic model on ROI *j*. The parameter δ_j controls the overall effect size on phenotype *j*, whereas $\zeta_{j1},\ldots,\zeta_{jk}$ are parameters controlling the contribution of each one of the *k* causative markers.

Compared to the average baseline GM value (calibrated on real data), we require the mean intensity value of the j^{th} affected ROI to be reduced by exactly $\gamma_j\times100\%$, where $\gamma_j\in[0,1]$ represents the overall genetic effect size. Therefore we impose that $E(y_j^*)=E(y_j)(1-\gamma_j)$ and solve for γ_j. The resulting expression,

$$\gamma_j = \frac{E(w_j)}{E(y_j)} = \frac{2\delta_j \sum_{s=1}^{k} \zeta_{js} m_s}{E(y_j)}$$

shows that the percentage reduction in GM at the j^{th} ROI depends on the mean baseline value, the observed MAF m_s for each causative SNP *s* $(s=1,\ldots,k)$ and the δ_j parameter $(j=1,\ldots,l)$. In our simulation settings, we control the effect size γ_j—since all other parameters are observed in the population, δ_j is then uniquely determined. We also

[4] http://www.loni.ucla.edu/ADNI.

[5] http://www.adni-info.org.

[6] http://www.fmrib.ox.ac.uk/fsl/fslview/atlas-descriptions.html.

Please cite this article as: Vounou, M., et al., Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.07.002

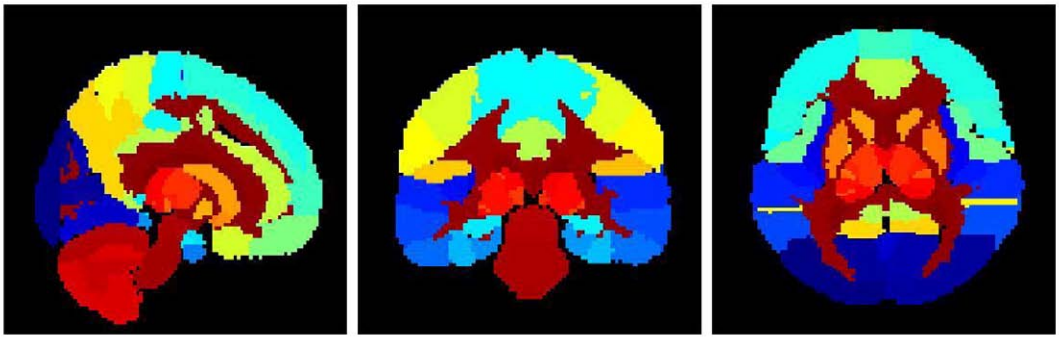

Fig. 1. Sagittal, coronal, and axial views of the GSK CIC Atlas defining 111 regions of interest.

report on the percentage of variance explained by the genetic effect for each phenotype j,

$$v_j = \frac{\mathrm{Var}(w_j)}{\mathrm{Var}(y_j) + \mathrm{Var}(w_j)}.$$

Assuming that all SNPs contribute equally, it can be noted that the effect on the mean GM of ROI j caused by a single causative SNP with MAF m is exactly $2\delta_j m / kE(y_j)$. When a randomly selected individual has maximal allele dosage at all k causative SNPs, γ_j takes its maximal value $2\delta_j / E(y_j)$.

2.1.4. Simulation parameter settings

In our simulations, we set $\zeta_{js} = 1/k$ to have each causal SNP affect each ROI equally. Effect sizes represented by the γ_j parameters were selected to introduce a 6%, 8%, and 10% reduction in mean GM in each affected ROI. The corresponding average proportions of variance explained by the genetic effects are 5%, 8%, and 12%, respectively. The maximally attainable per-SNP effects, observed when an individual is homozygous for the disease allele, are 3%, 4%, and 5%, respectively. These effect sizes were selected with reference to previous imaging genetics findings. For instance, Filippini et al. (2009) reported a 10% reduction in GM in homozygote ApoE $\varepsilon 4$ subjects relative to subjects with no $\varepsilon 4$ alleles (corresponding to our baseline GM values), and Joyner et al. (2009) reported a maximum genetic effect of 9.8%. Therefore, the genetic effect sizes chosen in our simulation studies are meant to characterise the statistical power when the per-SNP effects are relatively small and when multiple disease alleles contribute additively. Each simulation scenario consists of a unique parameter combination (γ, n) indicating the overall genetic effect size and sample size, respectively. To avoid biases introduced by random sampling, for each simulation scenario, we always report on average performance measures, where the average is taken over a total of $B = 200$ independent samples extracted from the population.

2.2. Sparse reduced-rank regression (sRRR)

Based on a random sample of size n, we denote by $\mathbf{X}$ the $n \times p$ design matrix of genetic markers, and by $\mathbf{Y}$ the associated $n \times q$ matrix of phenotypes, and assume $n \ll p$. We do not consider here additional non-genetic confounding variables though these could be easily accommodated. The standard multivariate multiple linear regression (MMLR) model is

$$\mathbf{Y} = \mathbf{XC} + \mathbf{E} \tag{1}$$

where $\mathbf{C}$ is the $(p \times q)$ matrix of regression coefficients and $\mathbf{E}$ is the $(n \times q)$ matrix of errors. If n were greater than p, $\mathbf{C}$ could be estimated by least squares as

$$\hat{\mathbf{C}}_{(R)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{2}$$

and $\hat{\mathbf{C}}_{(R)}$ would be full rank, $R = min(p,q)$.

Even under such an unrealistic assumption concerning the sample size, there would still be significant limitations. First, it is well known that little is gained by formulating the multivariate multiple regression in these terms, in the sense that the same solution can be obtained by performing q independent regressions, one for each univariate response (Hastie et al., 2001; Izenman, 2008). Thus, the unconstrained regression model (1) essentially makes no use of any structure that may exist in the multivariate response. Second, with high-dimensional genetic variables, which are often characterised by patterns of nonrandom associations, the model would also suffer from multicollinearity—the lack of orthogonality among the covariates—which will inflate the variance of the regression coefficients. Lastly, and perhaps most importantly, the identification of the most important covariates would need to rely exclusively on the statistical significance of the unconstrained regression coefficients, thus requiring to deal with the massive multiple testing problem. In realistic settings, when n never exceeds p, another major complication is created by the fact that $(\mathbf{X}'\mathbf{X})$ is noninvertible, and therefore, some form of regularisation is always needed.

A solution to the first two issues above consists in imposing a rank condition on the regression coefficient matrix, namely that rank($\mathbf{C}$) is $R^* \leq min(p,q)$, as in the reduced-rank regression (RRR) model (Reinsel and Velu, 1998). Reducing the rank leads to an effective decrease in the number of parameters that need to be estimated and enables to exploit the multivariate nature of the response. Our aim is to derive an estimation procedure such that the resulting coefficient matrix $\mathbf{C}$ has two important properties: (a) is it of reduced rank R^* and (b) it has zero-entries in both row and columns corresponding to all covariates (genotypes) and responses (phenotypes) that should be excluded from the model.

If $\mathbf{C}$ has rank r, with $r = 1, \ldots, R$, it can be written as a product of a $(p \times r)$ matrix $\mathbf{B}$ and $(r \times q)$ matrix $\mathbf{A}$, both of full rank, i.e. rank($\mathbf{A}$) = rank($\mathbf{B}$) = r. The RRR model is thus written

$$\mathbf{Y} = \mathbf{XBA} + \mathbf{E}, \tag{3}$$

For a fixed rank r, the matrices $\mathbf{A}$ and $\mathbf{B}$ are obtained by minimising the weighted least squares criterion

$$\mathrm{M} = \mathrm{Tr}\{(\mathbf{Y} - \mathbf{XBA})\Gamma(\mathbf{Y} - \mathbf{XBA})'\} \tag{4}$$

for a given $(q \times q)$ positive definite matrix Γ. Most commonly the weight matrix Γ is set to be either the inverse of the estimated covariance matrix of the responses or the identity matrix. As detailed in the Appendix, these choices of Γ reveal connections to other multivariate models. The estimates $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ that minimise (4) are obtained as

$$\hat{\mathbf{A}} = \mathbf{H}'\Gamma^{-\frac{1}{2}} \tag{5}$$

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\Gamma^{\frac{1}{2}}\mathbf{H}$$

where $\mathbf{H}$ is the $(q \times r)$ matrix whose columns are the first r normalized eigenvectors associated with the r largest eigenvalues of the $(q \times q)$ matrix

$$\mathbf{R} = \Gamma^{\frac{1}{2}}\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\Gamma^{\frac{1}{2}} \tag{6}$$

Moreover, $\hat{\mathbf{B}}$ can be rewritten in terms of the least squares solution of Eq. (2),

$$\hat{\mathbf{B}} = \hat{\mathbf{C}}_{(R)}\Gamma^{\frac{1}{2}}\mathbf{H}. \tag{7}$$

Thus, the rank r estimate of the RRR coefficient matrix $\mathbf{C}$ is

$$\hat{\mathbf{C}}_{(r)} = \hat{\mathbf{B}}\hat{\mathbf{A}} = \hat{\mathbf{C}}_{(R)}\Gamma^{\frac{1}{2}}\mathbf{H}\mathbf{H}'\Gamma^{-\frac{1}{2}} \tag{8}$$

As the solutions $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ depend on normalised eigenvectors, they must satisfy

$$\begin{aligned} \mathbf{A}\Gamma\mathbf{A}' &= \mathbf{I} \\ \mathbf{B}'\mathbf{X}'\mathbf{X}\mathbf{B} &= \Lambda^2 \end{aligned} \tag{9}$$

where Λ^2 is the $(r \times r)$ diagonal matrix with diagonal entries the eigenvalues corresponding to the r eigenvectors in $\mathbf{H}$.

This factorisation of the regression coefficient $\mathbf{C} = \mathbf{BA}$, enables us to apply separate sparsity constraints on each of $\mathbf{A}$ and $\mathbf{B}$ related to phenotype and genotype variable selection respectively. For instance, in CP-GWA studies, only sparsity in $\mathbf{B}$ will be required, whereas in BW-GWA studies, both $\mathbf{A}$ and $\mathbf{B}$ are required to be sparse.

In high-dimensional problems, when the number of variables in both domains greatly exceeds the number of observations, it is common to assume that the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$ are diagonal. In fact, this has been successfully done in studies involving genomic and gene expression data, also posing complex correlational structures (Parkhomenko et al., 2009; Witten et al., 2009). Taking this strategy, i.e. estimating $\mathbf{X}'\mathbf{X}$ by $\mathbf{I}_p$ and also setting Γ equal to $\mathbf{I}_q$, Eq. (4) can be rewritten as

$$\mathrm{M} = \mathrm{Tr}\{\mathbf{YY}'\} - 2\mathrm{Tr}\{\mathbf{AY}'\mathbf{XB}\} + \mathrm{Tr}\{\mathbf{AA}'\mathbf{B}'\mathbf{B}\}. \tag{10}$$

Noting that the first term does not depend on $\mathbf{A}$ or $\mathbf{B}$, a sparse rank-one model is obtained by solving the corresponding penalised least squares problem,

$$\mathop{\mathit{arg\,min}}_{\mathbf{a},\mathbf{b}}\{-2\mathbf{a}\mathbf{Y}'\mathbf{X}\mathbf{b} + \mathbf{a}\mathbf{a}'\mathbf{b}'\mathbf{b} + \lambda_{\mathbf{a}}\|\mathbf{a}'\|_1 + \lambda_{\mathbf{b}}\|\mathbf{b}\|_1\} \tag{11}$$

where an L1 penalty has been added to penalise both coefficients, $\mathbf{a}$ and $\mathbf{b}$. Constraining the norms of the coefficients results in estimates that are shrunk towards zero. In ridge and Lasso regression (Hoerl and Kennard, 1970; Tibshirani, 1996), constraints are imposed on the L2 and L1 norms of the coefficients, respectively. While an L2 penalty results in shrunken estimates that achieve stability over least squares estimates, it does not guarantee sparsity in the estimates. In contrast, penalising the L1 norm of the coefficients results in sparse estimates. The penalisation parameters λ_a and λ_b control the sparsity and hence the number of explanatory variables and responses that are included in the model. When both λ_a and λ_b are zero, no variable selection is performed.

Penalised regression with convex penalties can be efficiently solved using coordinate descent algorithms that iteratively update the coefficient estimates using soft-thresholding (Friedman et al., 2007). Similarly, our optimisation problem is biconvex in **a** and **b** and can be solved iteratively. For fixed $\mathbf{a}$ and fixed penalisation parameter $\lambda_{\mathbf{b}}$,

$$\begin{aligned} \hat{\mathbf{b}} &= \mathop{\mathit{arg\,min}}_{\mathbf{b}}\{-2\mathbf{a}\mathbf{Y}'\mathbf{X}\mathbf{b} + \mathbf{a}\mathbf{a}'\mathbf{b}'\mathbf{b} + \lambda_{\mathbf{b}}\|\mathbf{b}\|_1\} \\ &= \frac{1}{\mathbf{a}\mathbf{a}'}S_{\lambda_{\mathbf{b}}}(\mathbf{X}'\mathbf{Y}\mathbf{a}) \end{aligned} \tag{12}$$

where $S_\lambda(\mathbf{k}) = \mathrm{sign}(\mathbf{k})\left(|\mathbf{k}| - \frac{\lambda}{2}\right)_+$ is the soft thresholding operator and $(\cdot)_+ = \mathit{max}(\cdot, 0)$. For fixed $\mathbf{b}$ and $\lambda_{\mathbf{a}}$,

$$\begin{aligned} \hat{\mathbf{a}} &= \mathop{\mathit{arg\,min}}_{\mathbf{a}}\{-2\mathbf{a}\mathbf{Y}'\mathbf{X}\mathbf{b} + \mathbf{a}\mathbf{a}'\mathbf{b}'\mathbf{b} + \lambda_{\mathbf{a}}\|\mathbf{a}'\|_1\} \\ &= \frac{1}{\mathbf{b}'\mathbf{b}}S_{\lambda_{\mathbf{a}}}(\mathbf{b}'\mathbf{X}'\mathbf{Y}) \end{aligned} \tag{13}$$

Starting with initial arbitrary coefficient vectors $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$, the solutions are found by using the updates (12) and (13) iteratively until convergence, with normalization conditions (9) enforced after each iteration. A schematic illustration of both MMLR and sRRR models is given in Fig. 2.

After the rank-one sparse solution has been found, further ranks can be obtained from the residuals of the data matrices, $\mathbf{X}$ and $\mathbf{Y}$. Precisely, once the d^{th} pair of regression coefficients, $\hat{\mathbf{b}}_d$ and $\hat{\mathbf{a}}_d$, has been obtained, the vectors $\hat{\mathbf{z}}_d = \mathbf{X}\hat{\mathbf{b}}_d$ and $\hat{\mathbf{w}}_d = \mathbf{Y}\hat{\mathbf{a}}_d'$ are computed and the residual matrices are formed as $\mathbf{X}^* = \mathbf{X} - \hat{\gamma}\hat{\mathbf{z}}_d$ and $\mathbf{Y}^* = \mathbf{Y} - \hat{\delta}\hat{\mathbf{w}}_d$, where $\hat{\gamma}$ and $\hat{\delta}$ are obtained from regressing $\mathbf{X}$ on $\hat{\mathbf{z}}_d$ and $\mathbf{Y}$ on $\hat{\mathbf{w}}_d$.

2.3. The rank trace plot

The search for an 'optimal' reduced-rank R^* can be aided by the *rank trace* plot (Izenman, 2008). The principle behind this graphical procedure is that, when an adequate rank r has been selected, the estimated sRRR coefficient matrix, $\hat{\mathbf{C}}_{(r)}$, should be close to the full rank

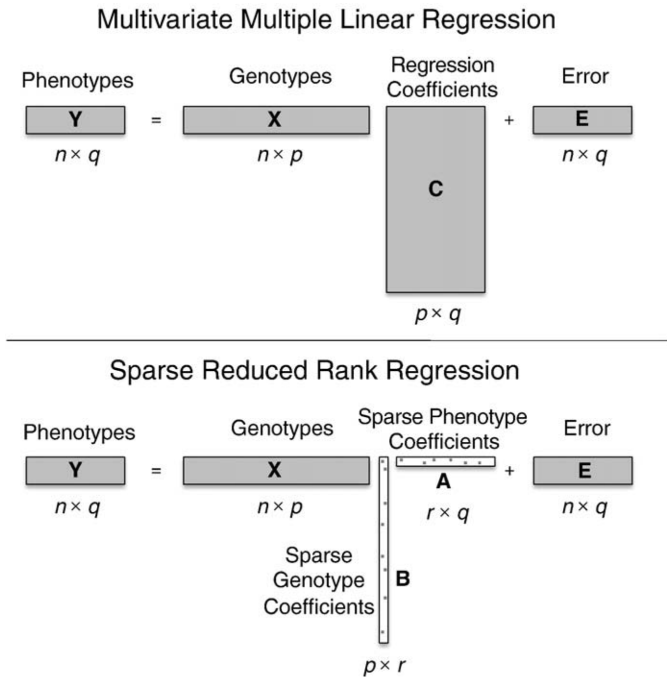

Fig. 2. Illustration of the multivariate multiple linear regression model and the sparse reduced rank regression model. Both are multivariate models, but the former cannot be fit unless sample size n exceeds p or constraints are placed on $\mathbf{C}$.

coefficient matrix $\hat{\mathbf{C}}_{(R)}$ and the estimated residual covariance matrix of the sRRR model,

$$\hat{\mathbf{S}}_{\epsilon\epsilon_{(r)}} = (\mathbf{Y}-\mathbf{X}\hat{\mathbf{C}}_{(r)})'(\mathbf{Y}-\mathbf{X}\hat{\mathbf{C}}_{(r)})$$

should be close to the corresponding full rank residual covariance $\hat{\mathbf{S}}_{\epsilon\epsilon_{(R)}}$. The rank trace is obtained by plotting, for all values of r in a range from 0 to R, the following two quantities:

$$\Delta\hat{\mathbf{C}}_{(r)} = \frac{\|\hat{\mathbf{C}}_{(R)}-\hat{\mathbf{C}}_{(r)}\|_F}{\|\hat{\mathbf{C}}_{(R)}\hat{\mathbf{C}}_{(0)}\|_F}$$

and

$$\Delta\hat{\mathbf{S}}_{\epsilon\epsilon_{(r)}} = \frac{\|\hat{\mathbf{S}}_{\epsilon\epsilon_{(R)}}-\hat{\mathbf{S}}_{\epsilon\epsilon_{(r)}}\|_F}{\|\hat{\mathbf{S}}_{\epsilon\epsilon_{(R)}}-\hat{\mathbf{S}}_{\epsilon\epsilon_{(0)}}\|_F}.$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The coefficient $\Delta\hat{\mathbf{C}}_{(r)}$ quantifies the relative change in the size of the regression coefficients between a rank r and the random model ($r=0$), holding the full rank model as reference. Similarly, the coefficient $\Delta\hat{\mathbf{S}}_{\epsilon\epsilon_{(r)}}$ represents the proportional difference in the corresponding residual covariance matrices. As r varies from 0 to R in both x and y axes, both coefficients take values in $[0,1]$. The two opposite points in the plot—those having coordinates $(0,0)$ and $(1,1)$—indicate the two extreme models: a full rank model ($r=R$) and a random model ($r=0$), respectively, where $\hat{\mathbf{C}}_{(0)}=0$ and $\hat{\mathbf{S}}_{\epsilon\epsilon_{(0)}}=\hat{\mathbf{S}}_{yy}$. As more ranks are added, starting at the top-right corner with $r=0$, the curve moves towards the origin of the plot. When a further rank addition does not produce a significant reduction in $\Delta\hat{\mathbf{C}}_{(r)}$ and $\Delta\hat{\mathbf{S}}_{\epsilon\epsilon_{(r)}}$, the plot indicates that an 'optimal' rank R has been found. In our experience, the rank corresponding to the point which maximises the curvature yields satisfactory results—this can be found by fitting a polynomial smoothing spline to the $(\Delta\hat{\mathbf{C}}_{(r)}, \Delta\hat{\mathbf{S}}_{\epsilon\epsilon_{(r)}})$ points for which second derivatives can be easily evaluated.

2.4. Performance assessment criteria

We evaluate the performance of sRRR, and compare it to MULM's performance, by means of ROC (receiver operating characteristic) curves. In each curve, sensitivity (true-positive rate) is plotted against 1 − specificity (false-positive rate) (Fawcett, 2004). This eschews multiple-testing correction or other model selection issues, as sensitivities can be compared for a given specificity. We separately evaluate the detection performance in genetic and imaging domains. In the sRRR, the "detected" SNPs correspond to all non-zero entries of $\hat{\mathbf{b}}_r$ ($r=1,\ldots,R^*$). As the penalty parameter λ_b is increased away from zero, sparser solutions are obtained and a smaller number of SNPs is retained. In MULM, SNPs are ordered in decreasing order of significance, according to the P value associated to each SNP–ROI pair. Since the true causative markers have been removed from the data, we define "true signal" SNPs as those that are LD-linked with at least one causal SNP. Specifically, any detected SNP whose R^2 coefficient with any of the causative SNPs is at least 0.8 is considered a true positive, with all others labelled as false positives. This LD threshold is commonly used in the literature, for example, for tagging SNPs (de Bakker et al., 2005; Wang et al., 2005). While the specific threshold may impact the absolute performance somewhat, the relative performance between statistical methods will be unaffected. We measure sensitivity as the proportion of true signal SNPs correctly detected, and false-positive rate as the proportion of true null SNPs incorrectly detected. Analogously for ROIs, sRRR selects a phenotype when its corresponding coefficient in $\hat{\mathbf{a}}_r$ ($r=1,\ldots,R$) has a nonzero element; the number of detected ROIs from MULM is then obtained accordingly from the ordered list of SNP–ROI pairs.

3. Results

The map of LD among the first 1000 available markers in the simulated population is represented in Fig. 3. The LD patterns resemble those observed in real human populations where neighbouring markers tend to be in high LD, and the pairwise LD coefficient between two markers decline with the distance between them. We report on simulation results obtained from subsets of the entire set of available markers, with the number of markers, p taking values of 1990, 9990, 19,990, and 37,738. Fig. 4 shows the number of LD-linked SNPs as a function of the LD threshold. Our threshold of 0.8 gives exactly 51 LD-linked SNPs, which correspond to approximately 2.56%, 0.51%, 0.26%, and 0.14% of the total number of SNPs, respectively, for the four values of p that we have considered.

Pairwise correlations among $q=111$ ROIs defined by the GSK CIC Atlas, estimated using 189 MCI subjects from the ADNI data set, are shown in Fig. 5. The inset shows the correlations among the 6 affected ROIs in the frontal cortex. The interregional correlations in the ADNI dataset were mostly positive, and strongest among cortical regions, with cerebellar and thalamic regions nearly independent of cortical regions.

When applying the sRRR model, a decision has to be made on how many ranks to select and how many variables to retain from each rank in both the genotype and phenotype spaces. In the statistical analysis of only one data set, these parameters would be optimally tuned using model selection criteria such as the cross-validated prediction error (see the Discussion and Appendix for further comments). In our simulation study, however, in which $B=200$ samples are extracted from the population for each given parameter setting, performing model selection is infeasible due to time and computation constraints. Guided by rank trace plots (see Fig. 11), we take the reduced-rank for all sRRR models to be $R^*=3$. However, the choice of how many SNPs and ROIs to retain from each one of the three ranks (i.e., how many zero coefficients to enforce in each $\mathbf{a}_r$ and $\mathbf{b}_r$, with $r=1,2,3$) is difficult. When $R^*=3$, a model selection procedure would provide the

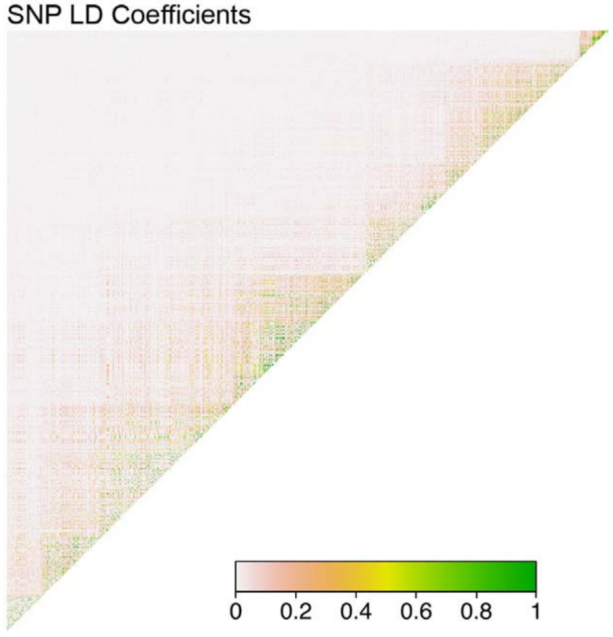

Fig. 3. A map of all pairwise LD coefficients for a subset of 1000 FREGENE-simulated SNPs used in this study. The simulated genetic data present the typical LD structure observed in real populations, where markers that are physically close to each other on the chromosome are in stronger LD, leading to a characteristic *block-like* structure.

Please cite this article as: Vounou, M., et al., Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.07.002

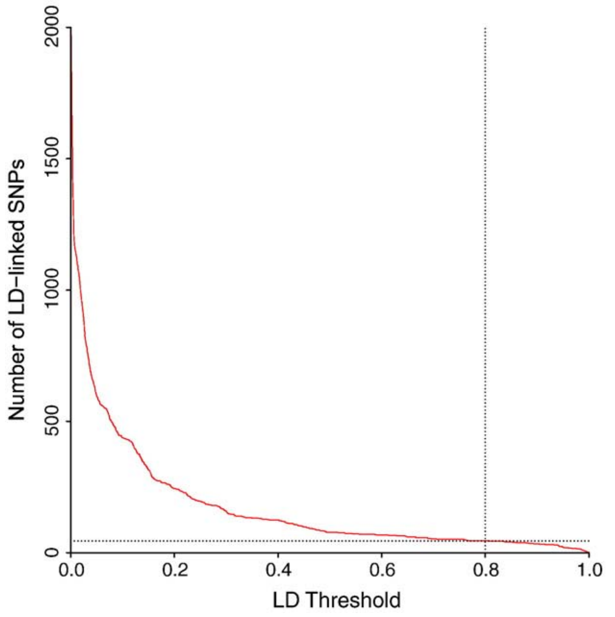

Fig. 4. Number of LD-linked SNPs (out of 1990 SNPs) as function of the LD threshold. Most SNPs have R^2 with causative SNPs that is 0.4 or less; only 51 SNPs with R^2 exceeding 0.8 were marker as “true” signal SNPs after the causal SNPs were removed from the analysis.

optimal allocation (h_1, h_2, h_3), meaning that $h_1 > 0$ variables are selected from the first rank, $h_2 > 0$ from the second, and $h_3 > 0$ from the third. For most results reported here, we have applied the simplest possible rule of uniform allocation across ranks: we vary the total number of variables to be retained, g, and use the allocation $(g/3, g/3, g/3)$, meaning that 1/3 of the g variables to be retained (either SNPs or ROIs) is selected from each rank. In some cases, we have tested the $(g-2, 1, 1)$ rule—we select all but two variables from the first rank, and then one variable for each one of the remaining two ranks. Although these allocations are arbitrary and do not guarantee that the sRRR model will always produce optimal ROC curves, they free us from the computational burden introduced by any data-intensive model selection procedure, thus allowing us to carry out an exhaustive exploration of several parameter combinations, including different effect sizes and sample sizes. Due to lack of optimisation, the results obtained using sRRR are conservative, and we expect that a full procedure that includes model selection will generally perform better.

Fig. 6 shows the ROC curves for SNP selection obtained from applying sRRR with three different reduced ranks $R^* = 1,2,3$ on $p = 1990$ SNPs and with a 6% effect size; the sample sizes are 500 (a) and 1000 (b), respectively. The corresponding ROC curves obtained from MULM are also shown for comparison. These curves show that sRRR demonstrates consistently better power than MULM for every level of specificity. As expected after inspection of the rank trace plots, when only one rank is used, not all LD-linked SNPs are detected by sRRR and thus MULM performs slightly better for some portions of the

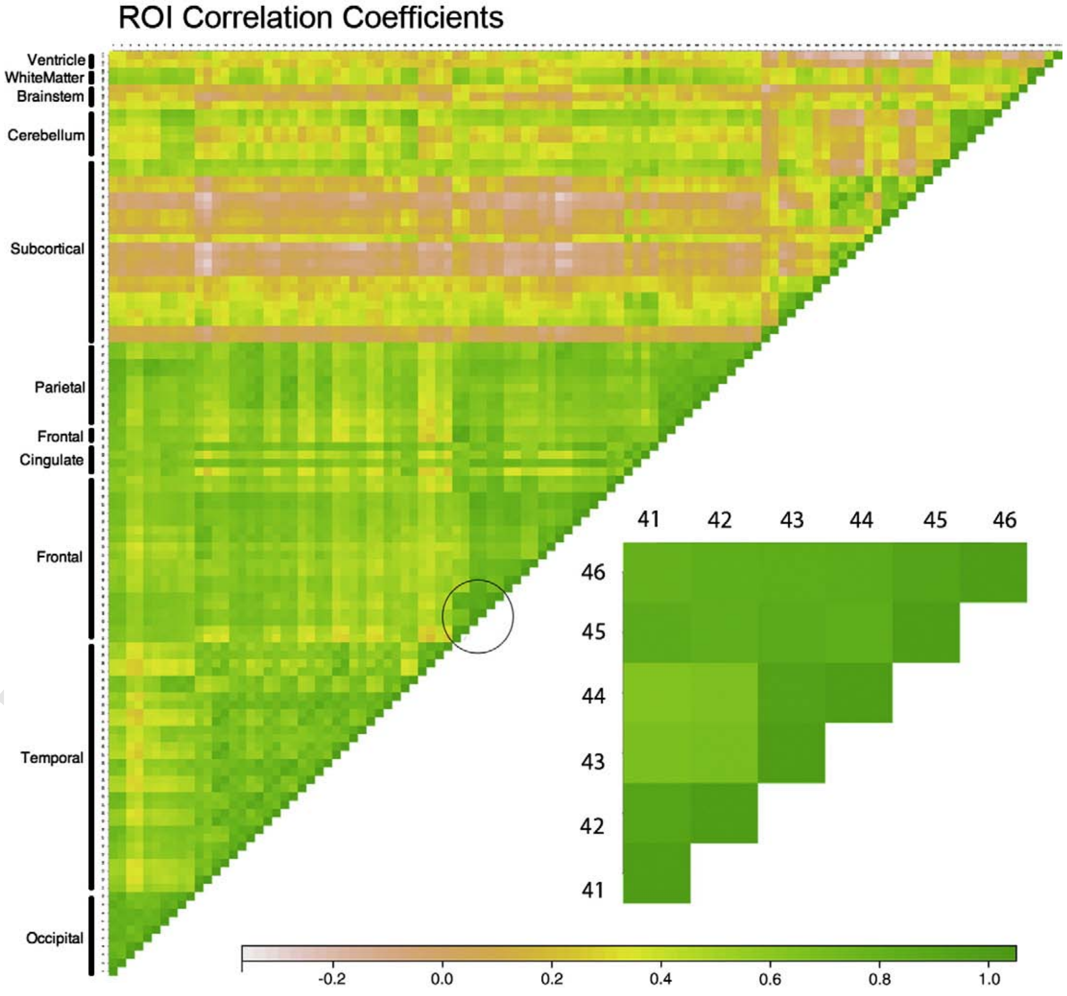

Fig. 5. All pairwise correlations among $q = 111$ ROIs defined by the GSK CIC Atlas and estimated using $n = 189$ MCI subjects from the ADNI data set. The inset shows the correlations among the 6 affected ROIs in the frontal cortex: left and right each of precentral gyrus (41, 42), anterior dorsolateral prefrontal cortex (43, 44), posterior dorsolateral prefrontal cortex (45, 46).

Please cite this article as: Vounou, M., et al., Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.07.002

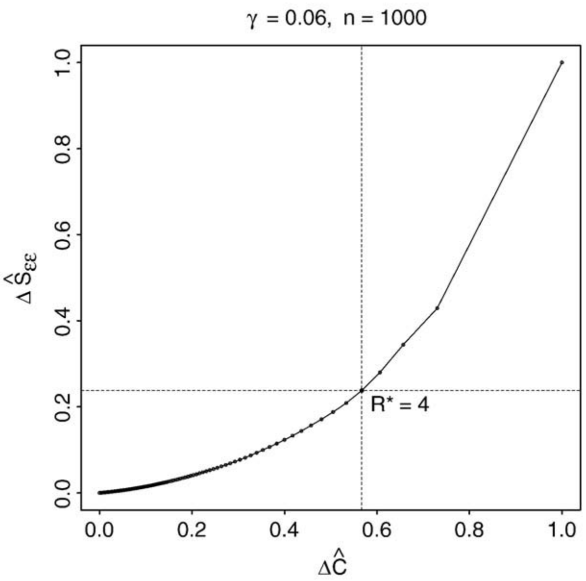

Fig. 6. ROC curves for SNP selection with a genetic effect size $\gamma=0.06$ and sample sizes $n=500$ (a) and $n=1000$ (b). The four ROC curves refer to sRRR with $R^*=1,2,3$ and to the mass-univariate approach based on several linear models (MULM). For almost all specificities considered, the sRRR method has always higher sensitivity than linear models—only when $n=500$ and $R^*=1$ the mass-univariate approach performs slightly better for some portions of the curve. The sensitivity of sRRR increases substantially when adding two ranks, and increases again when adding three ranks. All results are obtained as averages of $B=200$ replicates.

corresponding curve. In all cases, a notable gain in performance is obtained when increasing the rank from 1 to 2, with performances then improving marginally less as more ranks are added. This is in agreement with the rank trace plots and confirms that the true signal is captured by the first few ranks.

Fig. 7 shows the SNP detection performance when $R^*=3$ is used, with genetic effect size $\gamma=0.06$ and sample sizes $n=500$ (a) and $n=1000$ (b). Analogous ROC curves obtained for the higher effect size of 10% areshown in Fig. 8. In all cases, while power falls off appreciably for high specificity, the sRRR method always has better sensitivity. Interestingly, while the sensitivity of MULM improves as genetic effects and sample sizes increases, it only increases linearly with false positive rates. In contrast, as the signal gets stronger or the sample size gets larger, the performance of sRRR improves by a larger factor especially at lower levels of specificity—this can be appreciated by the higher curvature of the sRRR ROC curves. It is also important to remember that such high sensitivity is obtained despite no attempt being made to select the best sparsity parameters—for instance, even if sRRR was able to detect more than $g/3$ true positives in the first rank, these will be go undetected under the $(g/3,g/3,g/3)$ allocation rule.

To understand how the performance of sRRR scales from 1000s to 10s of 1000s of total SNPs, we computed sensitivity and false positive rates of sRRR and MULM for various values of p while equating g, the number of selected SNP between the two methods. Table 1 reports on our findings for a model with $\gamma=0.06$ and $n=1000$, where p ranges from 1990 to 37,738 and g ranges from 30 to 450. For every setting considered, sRRR has smaller false positive risk (0.60 to 0.95 that of MULM) and larger power (1.72 to 4.66 times greater than MULM). Remarkably, the relative power of sRRR compared to MULM gets larger as p increases, for any value of g, but particularly so for smaller values of g, when fewer SNPs are selected. For one setting, Fig. 9 illustrates that the power ratio increases with the number of SNPs considered, with sRRR's power increasing by a large factor when nearly 40 k markers are included. This provides reassurance that, in full-scale GWA studies, sRRR can achieve a much higher power than MULM, while keeping the false-positive rate at acceptable levels.

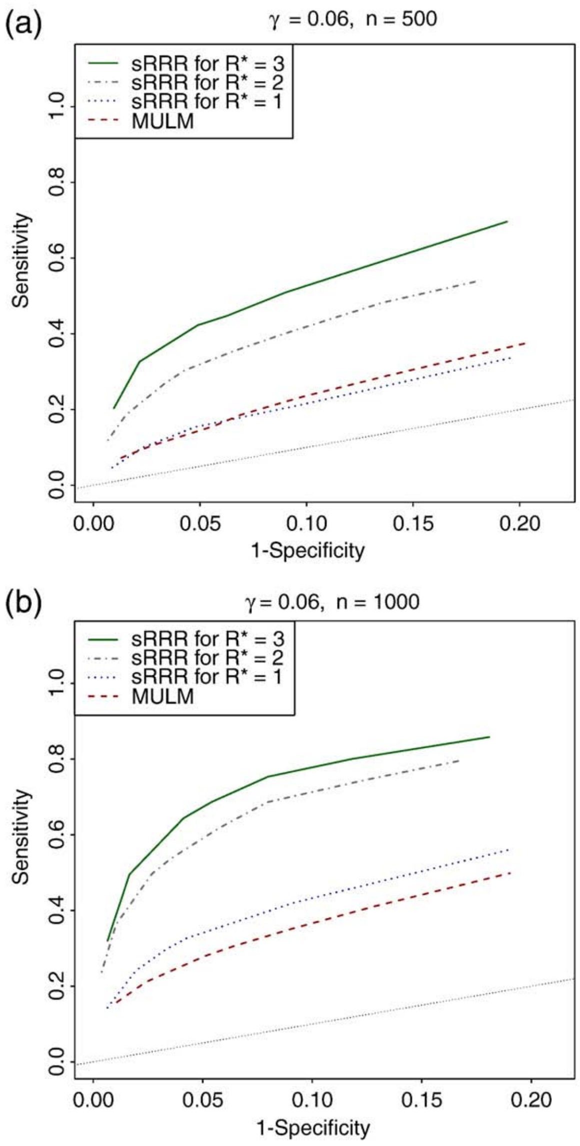

Fig. 7. ROC curves for SNP selection with a genetic effect size $\gamma=0.06$, $R^*=3$ selected ranks and sample sizes $n=500$ (a) and $n=1000$ (b). sRRR always outperforms mass-univariate linear models. With the sample size increases, the gain in sensitivity obtained from the mass-univariate approach is pretty much the same at all specificities, while the sRRR yield higher sensitivity corresponding to low specificity levels, which results in curves with higher curvature. All results are obtained as averages of $B=200$ replicates.

Under our simulated genetic effects, the power of either method rarely reaches the desired 80% this indicating the serious challenge of WB-GWA with even $n=1000$ subjects.

An assessment of the ROI selection performance using ROC curves is reported in Fig. 10 for effect sizes of 6% (a) and 10% (b), with a sample size of 500 subjects. In these figures, we illustrate the effect of the two allocation rules, uniform allocation, and the $(g-2,1,1)$ selecting most variables from rank 1. For the smaller effect size of $\gamma=0.06$, sRRR has higher sensitivity compared to MULM, at all specificity levels, and for both rules. However, the limitation of these arbitrary allocation rules is evident when a genetic effect size $\gamma=0,1$ is used, in plot (b). Clearly, sRRR is able to detect the most important ROIs from rank 1, and the rule $(g-2,1,1)$ provides high sensitivity at low specificity. However, since 2 ROIs also need to be selected from the second and third ranks, MULM outperforms sRRR at lower specificity in this instance. At a slightly higher specificity level, when all the affected ROIs have been selected, sRRR achieves better power.

Please cite this article as: Vounou, M., et al., Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.07.002

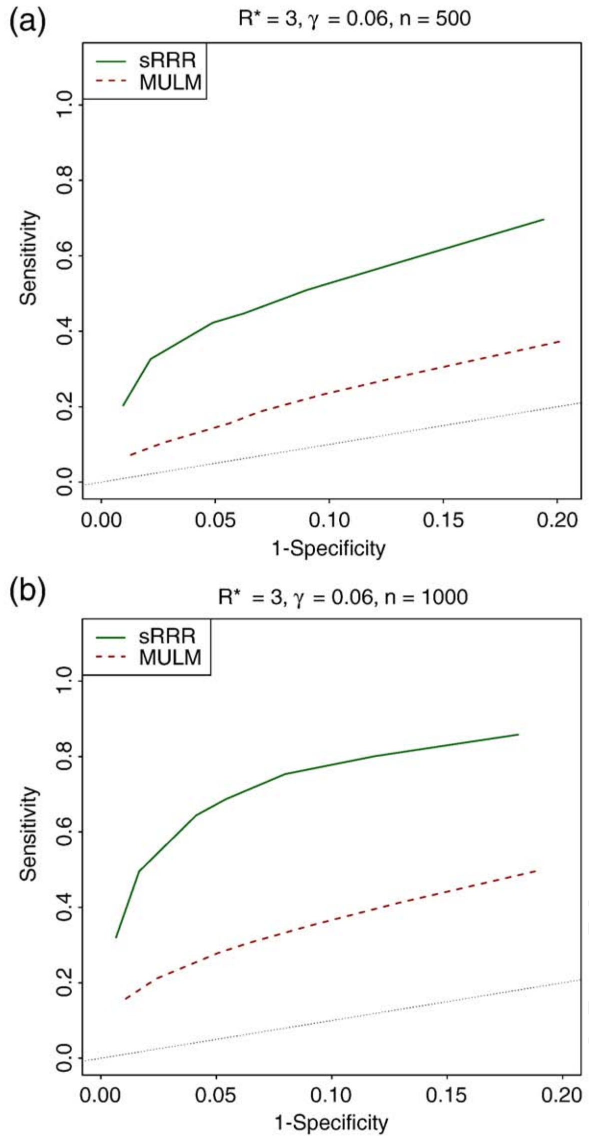

Fig. 8. ROC curves for SNP selection: genetic effect size $\gamma = 0.1$, $R^* = 3$ selected ranks and sample sizes $n = 500$ (a) and $n = 1000$ (b). sRRR always outperforms mass-univariate linear models. With the sample size increases, the gain in sensitivity obtained from the mass-univariate approach is pretty much the same at all specificities, while the sRRR yield higher sensitivity corresponding to low specificity levels, which results in curves with higher curvature. All results are obtained as averages of $B = 200$ replicates.

Table 1
False-positive rate and power comparisons: p is the total number of available SNPs; g is the target number of selected SNPs; α_{sRRR} and α_{MULM} is the false positive rate (1 − specificity) achieved by sRRR and MULM, respectively. π_{sRRR} and π_{MULM} is the power (sensitivity) achieved by sRRR and MULM, respectively. In sRRR, we set $R^* = 3$ and use the uniform allocation rule ($g/3, g/3, g/3$). Note that due to possible redundancies between the sets of $g/3$ SNPs selected from each rank, the actual number of 'unique' SNPs, selected over all ranks, is usually somewhat less than the target number g, illustrated in this table. For any value of g, as the total number ofSNPs in the study gets larger, the ratio $\alpha_{sRRR}/\alpha_{MULM}$ remains constant and always below 1, indicating that sRRR achieves smaller false positive rate, while the ratio π_{sRRR}/π_{MULM} is always above 1, indicating that sRRR achieves higher power. Remarkably, the relative power of sRRR compared to MULM gets larger as p increases, for any value of g, but particularly so for smaller values of g. The sample size is $n = 1000$ and the genetic effect is $\gamma = 0.06$. All results are obtained as averages of $B = 200$ replicates.

p	g	g/p	α_{sRRR}	α_{MULM}	$\alpha_{sRRR}/\alpha_{MULM}$	π_{sRRR}	π_{MULM}	π_{sRRR}/π_{MULM}
1990	30	0.0151	0.0065	0.0108	0.6044	0.3201	0.1577	2.0292
9990		0.0030	0.0015	0.0024	0.6199	0.2825	0.1054	2.6809
19,990		0.0015	0.0008	0.0012	0.6249	0.2683	0.0866	3.0997
37,738		0.0008	0.0004	0.0007	0.6117	0.2607	0.0640	4.0720
1990	60	0.0302	0.0165	0.0240	0.6887	0.4953	0.2110	2.3476
9990		0.0060	0.0036	0.0052	0.6986	0.4400	0.1363	3.2288
19,990		0.0030	0.0019	0.0026	0.7108	0.4098	0.1134	3.6128
37738		0.0016	0.0010	0.0014	0.7011	0.4019	0.0862	4.6633
1990	120	0.0603	0.0413	0.0509	0.8114	0.6439	0.2792	2.3062
9990		0.0120	0.0087	0.0106	0.8177	0.5581	0.1814	3.0773
19,990		0.0060	0.0045	0.0054	0.8236	0.5192	0.1452	3.5760
37,738		0.0032	0.0024	0.0029	0.8204	0.4968	0.1093	4.5444
1990	150	0.0754	0.0540	0.0640	0.8435	0.6865	0.3056	2.2464
9990		0.0150	0.0114	0.0134	0.8479	0.5946	0.1970	3.0189
19,990		0.0075	0.0058	0.0069	0.8466	0.5698	0.1563	3.6462
37,738		0.0040	0.0031	0.0037	0.8508	0.5266	0.1203	4.3773
1990	210	0.1055	0.0798	0.0904	0.8829	0.7533	0.3511	2.1458
9990		0.0210	0.0170	0.0191	0.8873	0.6431	0.2227	2.8873
19,990		0.0105	0.0086	0.0097	0.8864	0.6055	0.1749	3.4619
37,738		0.0056	0.0046	0.0052	0.8890	0.5620	0.1360	4.1327
1990	300	0.1508	0.1184	0.1286	0.9204	0.8005	0.4112	1.9468
9990		0.0300	0.0254	0.0276	0.9211	0.6754	0.2515	2.6858
19,990		0.0150	0.0129	0.0140	0.9200	0.6316	0.1949	3.2404
37,738		0.0079	0.0068	0.0074	0.9192	0.5982	0.1545	3.8718
1990	450	0.2261	0.1808	0.1902	0.9503	0.8580	0.4985	1.7211
9990		0.0450	0.0393	0.0414	0.9489	0.7054	0.2934	2.4039
19,990		0.0225	0.0200	0.0211	0.9459	0.6723	0.2258	2.9774
37,738		0.0119	0.0106	0.0113	0.9451	0.6351	0.1779	3.5691

The limitation of the ($g/3, g/3, g/3$) allocation is also clearly demonstrated here—although sRRR achieves very high sensitivity and essentially detects all the affected ROIs with a false discovery rate of about 10%, it has low power at lower specificity, because only 1/3 of all total g variables can enter the model for each rank.

4. Discussion

We have tackled the problem of detecting associations between high-dimensional genetic and imaging variables by casting it as a multivariate regression problem with multiple responses. The traditional approach to multivariate regression is to estimate the coefficients by ordinary least squares and to use the resulting estimates for prediction. When the number of explanatory variables is large and many of them are highly correlated with each other, we demonstrate that it is advantageous to predict the responses with fewer linear combinations of the genetic explanatory variables. In our proposed reduced-rank regression, the predictions are obtained from a subspace of the space spanned by the explanatory variables.

An essential ingredient in our formulation is provided by the sparsity constraints, which effectively allow us to select highly predictive genetic markers. When thousands of markers are included in the model as potential casual variants (for instance, in GWA studies), the large majority of them is not expected to be involved with the disease under study. As a consequence, the underlying true, but unknown, regression model it necessarily thought of as being sparse: only a few markers, if any at all, have anon-zero regression coefficient, whereas the majority of them have no influence on the quantitative traits, and do not enter the model. Our proposed estimation procedure builds on these assumptions and produces sparse solutions accordingly. Sparsity at the phenotypic level is also required when the number of candidate quantitative traits entering the regression model is very large; for instance, when there are several candidate ROIs (as in our simulation setting) or in whole-brain analyses carried out at the voxel-level. In these cases, it is not known with certainty which quantitative phenotypes provide a good proxy for the disease, and the sRRR model is able to discover them alongside the casual genetic markers.

Our approach is related to other multivariate models that have been used to explore linear and nonlinear dependences between high-dimensional covariates and responses in a least squares framework, such as canonical correlation analysis (CCA) and partial least squares (PLS). These belong to a larger class of *latent variable*

Please cite this article as: Vounou, M., et al., Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.07.002

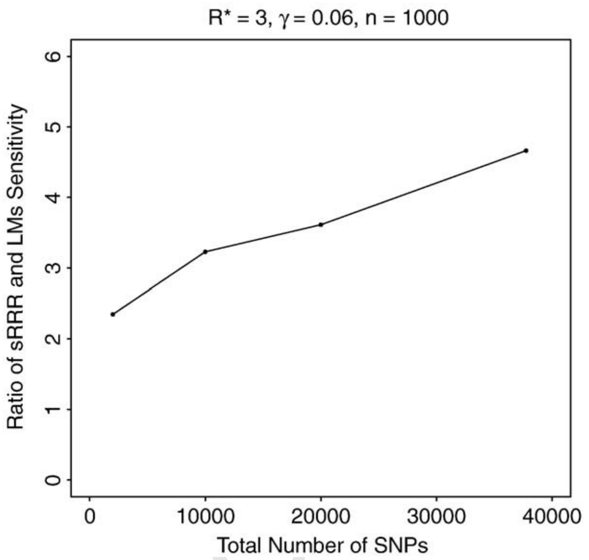

Fig. 9. Comparison of sRRR and MULM for large p: shown here is the ratio of SNP sensitivities (sRRR/LMs) as a function of the total number of SNPs included in the study. The genetic effect size is $\gamma = 0.06$, $R^* = 3$ selected ranks and sample size $n = 1000$. All results are obtained as averages of $B = 200$ replicates. This result suggests that the potential power gain coming from the sRRR model can be much higher in genome-wide scans when the number of available SNPs is much higher than 40 k. See Table 1 for further details.

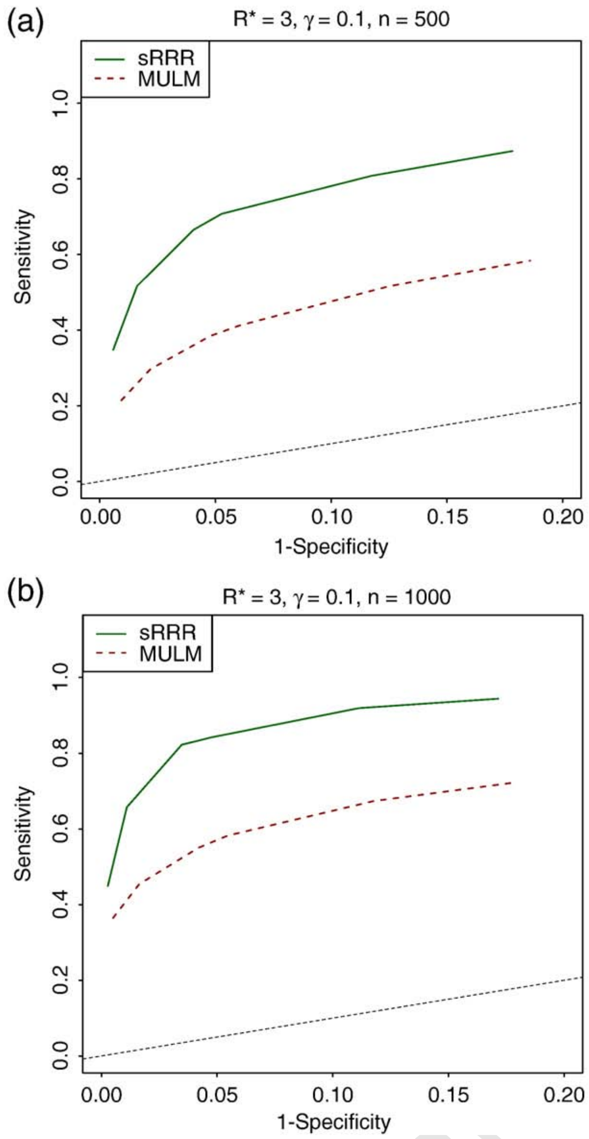

Fig. 10. ROC curves for ROI selection: $n = 500$, $R^* = 3$ and genetic effect size $\gamma = 0.06$ (a) and $\gamma = 0.1$ (b). For the latter genetic effect, sRRR method has worse specificity for lowest false-positive rates, and the mass-univariate approach shows good performance. Notably, for the lower genetic effect, sRRR outperforms linear models. The mass-univariate approach is expected to perform well in this task because all the affected ROIs are observed. All results are obtained as averages of $B = 200$ replicates.

models (LVMs) that perform dimensionality reduction in meaningful, albeit different, ways. When no response variables are available, other common examples of LVMs include PCA (principal component analysis) and ICA (independent component analysis). PCA extracts a handful of latent variables or *principal components* that explain as much sample variance as possible, while ICA seeks linear combinations of variables satisfying some optimal properties subject to mutual independence. Where two paired sets of variables are available, CCA finds *canonical variables* that explain as much sample correlation as possible between the two domains. Our proposed RRR model is closely related to both CCA and PLS (see Appendix).

In the analysis of genetic data, statistical models that assume the existence of some underlying hidden variables or latent *factors* having some optimal properties (such as maximal variance) have recently gained popularity. These approaches offer practical ways to deal with the widespread correlation patterns seen in genomic data and yield interpretable results. For instance, it has been observed that the first few principal components extracted from genetic markers capture the ancestral information contained in the sample and aid in the identification of population sub-structure (Reich et al., 2008). PCA also has a precise genealogical interpretation (McVean, 2009) and has been used for detecting *tagging* SNPs (Lin and Altman, 2004)—'landmark' markers that capture much variability in a given chromosomal region and can be used in place of many other neighbouring markers in an effort to reduce dimensionality. In case–control association studies, LVMs such as principal component regression (Wang and Abbott, 2008), ICA (Dawy et al., 2005) and PLS (Sarkis et al., 2006) have also been proposed to exploit correlations among SNPs.

In the analysis of imaging data, LVMs have been used widely, for instance in the modelling of correlation patterns and detecting dependences among brain regions. For instance, CCA has been used for the segmentation of magnetic resonance spectroscopic images (Laudadio et al., 2005), to estimate the shapes of obscured anatomical sections of the brain from visible structures in MRI (Liu et al., 2004) and to extract highly correlated modes of variation in shape between a number of different anatomical structures within the brain (Rao et al., 2006). In functional MRI studies, CCA has been proposed to identify activations of low contrast in the brain—by accounting for neighbouring correlated voxels, these models yield increased sensitivity to detect true signals relative to single voxel analyses (Friman et al., 2001; Nandy and Cordes, 2003). RRR with regularised covariance matrices has also been used as a predictive model of brain activation (Kustra, 2006).

Within the emerging field of *imaging genetics*, LVMs have only recently made their first appearance. A nonlinear extension of CCA, kernel CCA, has been used to investigate the association between a set of candidate SNPs and a set of voxels taken from the entire brain image (Hardoon et al., 2009)—in practice, a linear kernel was used, corresponding to a standard linear CCA. An extension of ICA that computes a dependence measure between two paired sets of variables, called parallel ICA (pICA), has also been proposed for imaging genetics studies. In pICA, latent variables are extracted by maximising the between-domain correlation while ensuring that all the extracted variables are as independent as possible within each domain (Liu et al., 2008). Both kernel CCA and pICA find shared

Please cite this article as: Vounou, M., et al., Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.07.002

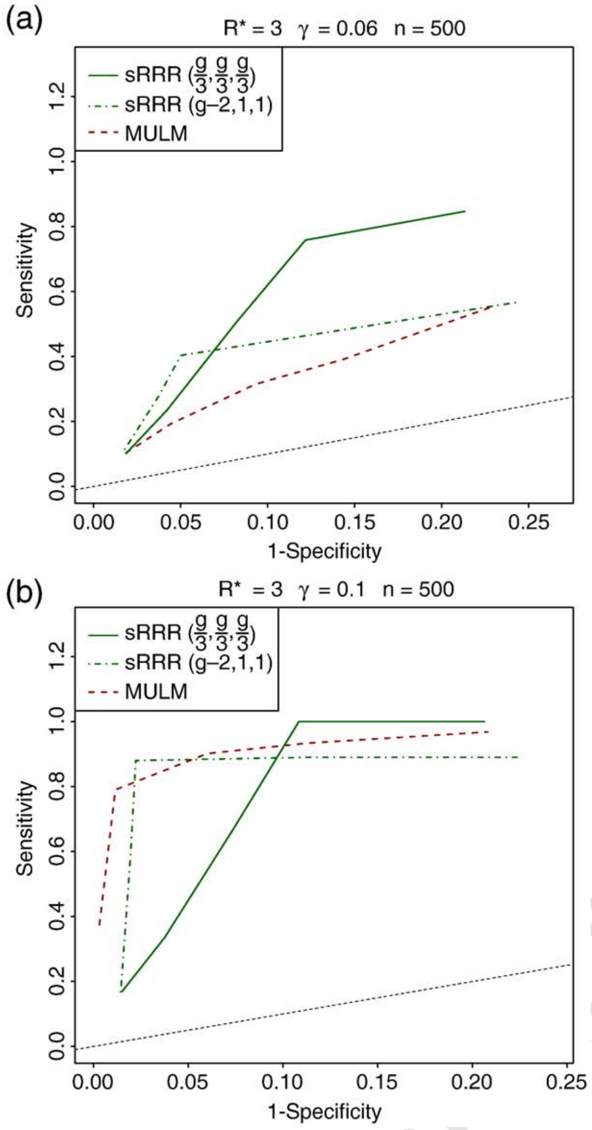

Fig. 11. Rank trace plot. In the *x*-axis, $\Delta\hat{C}$ is the ratio of two quantities: the difference between the regression coefficients obtained from a model with full rank and one with reduced-rank *r*, and the difference between the regression coefficients obtained from a model with full rank and a random model; in the *y*-axis, $\Delta\hat{S}_{\epsilon\epsilon}$ is the proportional difference in the corresponding residual covariance matrices. For each reduced-rank *r* ranging from 0 (top-right corner) to *R* (bottom-left corner), there is a corresponding point $(\Delta\hat{C}_{(r)}, \Delta\hat{S}_{\epsilon\epsilon_{(r)}})$ along the curve. A suitable rank R^* can be selected by locating the point at which curvature is maximal—in this example, based on $\gamma = 0.06$ and $n = 1000$, this point corresponds to $R = 4$ and is marked by the vertical and horizontal lines.

hidden factors that may explain the dependence between genetic and imaging variables. The underlying assumption is that such common factors are surrogates of the disease. However, the lack of sparsity in the solutions found by these models makes their interpretation particularly difficult as there are no rigorous criteria to rank genotypes and phenotypes by importance. Our model provides a solution to this problem by performing simultaneous variable selection in both domains in a *predictive modelling* fashion. We believe that the emphasis on variable selection is particularly important when the underlying (and unobserved) model that generated the 'true' association has a sparse representation, which is precisely the case in association mapping.

As already highlighted, the introduction of sparsity constraints raises important model selection issues that adds to the necessity of determining an adequate reduced rank—a task analogous to choosing the number of latent factors in CCA and pICA. On real datasets the selection of R^* can be accomplished by graphical devices such as the rank trace plot, which we find to perform well in practice. Permutation-based procedures, cross-validation, and parametric test statistics have also been proposed in similar problems (Reinsel and Velu, 1998; Waaijenborg et al., 2008; Witten et al., 2009). The issue of selecting the penalty coefficients that control how many variables in each domain enter the regression model can be addressed by adopting the cross-validated prediction error as a search criterion to be minimised (see Appendix for further details). We are currently developing analytical expressions for evaluating the cross-validated predictive performance of the sRRR model, thus allowing model selection to be performed quickly on very large data sets.

5. Conclusion

We have proposed a novel multivariate method, sparse reduced-rank regression (sRRR), for identifying associations between imaging phenotypes and genetic markers, and have performed detailed, calibrated simulations to evaluate its performance. Our results indicate that sRRR is a very promising approach and has high power to detect the most important variables in both the genetic and imaging domains. This is particularly the case at small sample sizes and with small genetic effects, where our method compares very favourably with more traditional univariate approaches. When increasing the number of genetic markers, the relative power obtained from sRRR compared to MULM increases with lower signal-to-noise ratios. This result further encourages the use of sRRR as an alternative procedure especially in the extremely high-dimensional BW-GWA paradigm. To the best of our knowledge, this is also the first assessment of statistical power in imaging genetics, and the first such comparison between univariate and multivariate methods.

Further work is currently under way to extend the proposed model in a number of directions including the implementation of alternative penalty functions, and to enable the detection of associations with markers in biological pathways, rather than individual markers. Our simulation framework could also be used to directly compare the power of traditional GWA studies, using only the case–control status as response, with that of BW-GWA studies that rely on multivariate responses.

6. Uncited reference Q1

Wang & Abbott, 2008

Acknowledgments

Maria Vounou is supported by a grant from the EPSRC and GlaxoSmithKline Clinical Imaging Center. The authors thank Matt Silver and Becky Inkster for comments on earlier versions of the manuscript. Imaging data were provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc., F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., and Wyeth, as well as nonprofit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by

Please cite this article as: Vounou, M., et al., Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.07.002

the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514, and the Dana Foundation.

Appendix

Connection of sRRR to latent variable models

The RRR model is closely related to two well-known multivariate dimensionality reduction methods: canonical correlation analysis (CCA) and partial least squares (PLS). Both models can be shown to be special cases of RRR for different choices of the matrix *Γ*. In this appendix, we briefly describe these models and clarify their connection with RRR.

CCA is a well known multivariate technique that reduces the dimensionality of the paired sets of variables by extracting $R^* \leq min(p,q)$, mutually orthogonal pairs of latent variables. These are formed as $\mathbf{T} = \mathbf{XU}$ and $\mathbf{S} = \mathbf{YV}$ where $\mathbf{U}$ and $\mathbf{V}$ are the $(p \times R^*)$ and $(q \times R^*)$ matrices of weights. Each pair of weight vectors $(\mathbf{u}_r, \mathbf{v}_r)$, $r = 1,\ldots,R^*$, forming the r^{th} columns of $\mathbf{U}$ and $\mathbf{V}$, is obtained so as to produce pairs of maximally correlated latent variables $\mathbf{t}_r = \mathbf{Xu}_r$ and $\mathbf{s}_r = \mathbf{Yv}_r$ that are orthogonal to the previously extracted latent variable pairs. The solutions $\mathbf{u}_r$ and $\mathbf{v}_r$ are extracted by maximising the correlation between $\mathbf{t}_r$ and $\mathbf{s}_r$, the so-called canonical correlation, given by

$$\mathrm{Corr}(\mathbf{t}_r, \mathbf{s}_r) = \frac{\mathbf{u'}_r \mathbf{X'Y}\mathbf{v}_\mathrm{r}}{\sqrt{\mathbf{u}'_r \mathbf{X}'\mathbf{X}\mathbf{u}_r \mathbf{v}'_r \mathbf{Y}'\mathbf{Y}\mathbf{v}_r}} \quad \text{for } r = 1,\ldots,R^*.$$

Unique solution are given by solving

$$\underset{\mathbf{u}_r,\mathbf{v}_r}{max}\, \mathbf{u}'_r \mathbf{X}'\mathbf{Y}\mathbf{v}_r \quad \text{such that} \quad \mathbf{u}'_r \mathbf{X}'\mathbf{X}\mathbf{u}_r = \mathbf{v}'_r \mathbf{Y}'\mathbf{Y}\mathbf{v}_r = 1$$

The weights for the first R^* CCA latent variables solve to

$$\mathbf{U} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-\frac{1}{2}}\mathbf{H}^*\mathbf{\Xi}^{-1}$$
$$\mathbf{V} = (\mathbf{Y}'\mathbf{Y})^{-\frac{1}{2}}\mathbf{H}^*$$

where $\mathbf{H}^*$ is the $(q \times R^*)$ matrix whose columns are the first R^* normalised eigenvectors of $\mathbf{R}^*$, where

$$\mathbf{R}^* = (\mathbf{Y}'\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-\frac{1}{2}} \tag{14}$$

and $\mathbf{\Xi}$ is a diagonal matrix composed of the square roots of the corresponding R^* eigenvalues; these coefficients are also equal to the canonical correlations of the R^* latent variable pairs. There is a close connection between the solutions of RRR and CCA. When *Γ* is set to be proportional to the inverse of the covariance of the responses, estimated as $(\mathbf{Y}'\mathbf{Y})^{-1}$, the $(q \times q)$ matrix $\mathbf{R}$ in Eq. (6) becomes identical to $\mathbf{R}^*$ in Eq. (14). Consequently, the matrix of weights $\mathbf{U}$ forms a scaled version of $\hat{\mathbf{B}}$, defined for RRR in Eq. (5). The scaling on each column of *fB* is a result of the different normalisation constraints imposed on each optimisation problem. Moreover, the matrix of weights $\mathbf{V}$ can be seen as a generalised inverse of $\hat{\mathbf{A}}$ defined for RRR in Eq. (5). Various estimation algorithms for obtaining sparse CCA solutions have been proposed (Lykou and Whittaker, 2009; Parkhomenko et al., 2009; Waaijenborg et al., 2008; Witten et al., 2009).

PLS is another widely used multivariate dimensionality reduction technique that finds pairs of latent variables $(\mathbf{t}_r, \mathbf{s}_r)$ having maximum covariance. Precisely, $\mathbf{u}_r$ and $\mathbf{v}_r$ are extracted by maximising

$$\mathrm{Cov}(\mathbf{t}_r, \mathbf{s}_r) = \mathbf{u}'_r \mathbf{X}'\mathbf{Y}\mathbf{v}_r \quad \text{such that} \quad \mathbf{u}'_r \mathbf{u}_r = \mathbf{v}'_r \mathbf{v}_r = 1$$

It can be noted that, due to the following covariance decomposition

$$\mathrm{Cov}(\mathbf{Xu}_r, \mathbf{Yv}_r)^2 = \mathrm{Corr}(\mathbf{Xu}_r, \mathbf{Yv}_r)^2 \mathrm{Var}(\mathbf{Xu}_r)\mathrm{Var}(\mathbf{Yv}_r)$$

the maximisation of sample variance explained by the latent factors also maximises the sample correlation between factors when the variance explained by each individual component is also maximised. The PLS solution for the first *R* latent variablesis given by

$$\mathbf{U} = \mathbf{X}'\mathbf{Y}\mathbf{H}^+\mathbf{M}^{-1}$$
$$\mathbf{V} = \mathbf{H}^+$$

where $\mathbf{H}^+$ is the $(q \times R^*)$ matrix whose columns are the first R^* normalised eigenvectors of $\mathbf{R}^+$, with

$$\mathbf{R}^+ = \mathbf{Y}'\mathbf{X}\mathbf{X}'\mathbf{Y} \tag{15}$$

The diagonal matrix $\mathbf{M}$ has entries given by the square roots of the R^* largest eigenvalues of $\mathbf{R}^+$ which equal to the covariances of the R^* latent variable pairs. Notably, CCA solutions also solve the PLS problem when the estimated covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$ are diagonal matrices. The same connection holds between RRR and PLS when additionally *Γ* is set to be the identity matrix. Alternative algorithms to obtain sparse PLS solutions have recently been derived (Chun and Keles, 2007; Le Cao et al., 2008).

Sparsity selection using cross-validated prediction error

The sparsity parameters (λ_a, λ_b) can be chosen so as to optimise a model selection criterion. Among other choices, one such criterion can be the cross-validated prediction error (CVPE), a measure of out-of-sample prediction accuracy that avoids overfitting. Holding the (λ_a, λ_b) pair fixed to some values, a full *K*-fold cross-validation procedure can be performed as follows. Assuming a random sample with *n* subjects, the sample is partitioned into two disjoint subsets called *training* and *testing* sets, with the testing set having approximately n/K subjects—there are *K* possible such sets. For each testing set, the sRRR model is fitted using the corresponding training set, that is data matrices $\mathbf{Y}^{[-k]}$ and $\mathbf{X}^{[-k]}$ $(k = 1,\ldots,K)$ obtained by removing all rows corresponding to subjects in the testing set. The model fit provides sparse estimates $\hat{\mathbf{a}}^{[-k]}$ and $\hat{\mathbf{b}}^{[-k]}$ or, when more than one rank is required, matrices $\hat{\mathbf{A}}^{[-k]}$ and $\hat{\mathbf{B}}^{[-k]}$. The procedure is then repeated by cycling through all *K* training and testing sets and the CVPE is computed as

$$\mathrm{CVPE} = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{nq} \|\mathbf{Y}^{[k]} - \mathbf{X}^{[k]}\hat{\mathbf{B}}^{[-k]}\hat{\mathbf{A}}^{[-k]}\|_F^2$$

where $\|\cdot\|_F^2$ is the square of the Frobenius norm. A search algorithm can be implemented to find the pair $(\hat{\lambda}_a, \hat{\lambda}_b)$ that minimises the CVPE.

References

Cantor, R.M., Lange, K., Sinsheimer, J.S., 2010. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Journal of Human Genetics 6–22.

Chun, H., Keles, S., 2007. Sparse partial least squares regression with an application to genome scale transcription factor analysis. Department of Statistics, University of Wisconsin, Madison, USA, Technical report.

Croiseau, P., Cordell, H.J., 2009. Analysis of North American Rheumatoid Arthritis Consortium data using a penalized logistic regression approach. BMC Proceedings 5, 1–5.

Please cite this article as: Vounou, M., et al., Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.07.002

Dawy, Z., Sarkis, M., Hagenauer, J., Mueller, J., 2005. A novel gene mapping algorithm based on independent component analysis. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings (ICASSP'05), 5.

de Bakker, P., Yelensky, R., Pe'er, I., Gabriel, S., Daly, M., Altshuler, D., 2005. Efficiency and power in genetic association studies. Nature genetics 37 (11), 1217–1223.

Donnelly, P., 2008. Progress and challenges in genome-wide association studies in humans. Nature 456 (7223), 728–731.

Fawcett, T., 2004. ROC graphs: notes and practical considerations for researchers. Machine Learning, p. 31.

Ferreira, M., Purcell, S., 2009. A multivariate test of association. Bioinformatics 25 (1), 132.

Filippini, N., Rao, A., Wetten, S., Gibson, R.A., Borrie, M., Guzman, D., Kertesz, A., Loy-English, I., Williams, J., Nichols, T., Whitcher, B., Matthews, P.M., 2009. Anatomically-distinct genetic associations of APOE epsilon4 allele load with regional cortical atrophy in Alzheimer's disease. NeuroImage 44 (3), 724–728.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., 2007. Pathwise coordinate optimization. Annals of Applied Statistics 1 (2), 302–332.

Friman, O., Cedefamn, J., Lundberg, P., Borga, M., Knutsson, H., 2001. Detection of neural activity in functional MRI using canonical correlation analysis. Magnetic Resonance in Medicine 45 (2), 323–330.

Good, C., Johnsrude, I., Ashburner, J., Henson, R., Friston, K., Frackowiak, R., 2001. A voxel-based morphometric study of ageing in 465 Normal adult human brains. Neuroimage 14 (1), 21–36.

Hardoon, D.R., Ettinger, U., Mourão Miranda, J., Antonova, E., Collier, D., Kumari, V., Williams, S.C.R., Brammer, M., 2009. Correlation-based multivariate analysis of genetic influence on brain volume. Neuroscience letters 450 (3), 281–286.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.

Hoerl, A., Kennard, R., 1970. Ridge regression: applications to nonorthogonal problems. Technometrics 12 (1), 69–82.

Hoggart, C., Whittaker, J., De Iorio, M., Balding, D., 2008. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genet 4 (7), e1000130.

Hoggart, C.J., Chadeau-Hyam, M., Clark, T.G., Lampariello, R., Whittaker, J.C., De Iorio, M., Balding, D.J., 2007. Sequence-level population simulations over large genomic regions. Genetics 177 (3), 1725–1731.

Izenman, A., 2008. Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer.

Joyner, A.H., Roddey, J.C., Bloss, C.S., Bakken, T.E., Rimol, L.M., Melle, I., Agartz, I., Djurovic, S., Topol, E.J., Schork, N.J., Andreassen, O.A., Dale, A.M., 2009. A common MECP2 haplotype associates with reduced cortical surface area in humans in two independent populations. PNAS 106 (36), 15475–15480.

Kustra, R., 2006. Reduced-rank regularized multivariate model for high-dimensional data. Journal of Computational and Graphical Statistics 15 (2), 312–318.

Kwee, L., Liu, D., Lin, X., Ghosh, D., Epstein, M., 2008. A powerful and flexible multilocus association test for quantitative traits. The American Journal of Human Genetics 82 (2), 386 Äì397.

Laudadio, T., Pels, P., De Lathauwer, L., Van Hecke, P., Van Huffel, S., 2005. Tissue segmentation and classification of MRSI data using canonical correlation analysis. Magnetic resonance in medicine: official journal of the Society of Magnetic Resonance in Medicine/Society of Magnetic Resonance in Medicine 54 (6), 1519–1529.

Le Cao, K., Rossouw, D., Robert-Granie, C., Besse, P., 2008. A sparse PLS for variable selection when integrating omics data. Statistical Applications in Genetics and Molecular Biology 7 (1), 35.

Lin, Z., Altman, R., 2004. Finding haplotype tagging SNPs by use of principal components analysis. The American Journal of Human Genetics 75 (5), 850–861.

Liu, J., Demirci, O., Calhoun, V., 2008. A parallel independent component analysis approach to investigate genomic influence on brain function. IEEE Signal Processing Letters 15, 413–416.

Liu, T., Shen, D., Davatzikos, C., 2004. Predictive modeling of anatomic structures using canonical correlation analysis. IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2004, pp. 1279–1282.

Lykou, A., Whittaker, J., 2009. Sparse CCA using a Lasso with positivity constraints. Computational Statistics and Data Analysis. .

McVean, G., 2009. A genealogical interpretation of principal components analysis. PLoS genetics 5 (10), e1000686.

Nandy, R., Cordes, D., 2003. Novel nonparametric approach to canonical correlation analysis with applications to low CNR functional MRI data. Magnetic Resonance in Medicine 50 (2).

Parkhomenko, E., Tritchler, D., Beyene, J., 2009. Sparse canonical correlation analysis with application to genomic data integration. Statistical Applications in Genetics and Molecular Biology 8 (1), 1.

Potkin, S., Turner, J., Guffanti, G., Lakatos, A., Fallon, J., Nguyen, D., Mathalon, D., Ford, J., Lauriello, J., Macciardi, F., et al., 2008. A genome-wide association study of schizophrenia using brain activation as a quantitative phenotype. Schizophrenia Bulletin.

Rao, A., Babalola, K., Rueckert, D., 2006. Canonical correlation analysis of sub-cortical brain structures using non-rigid registration. Lecture Notes in Computer Science 4057, 66–74.

Reich, D., Price, A.L., Patterson, N., 2008. Principal component analysis of genetic data. Nature Genetics 40 (5), 491–492.

Reinsel, G., Velu, R., 1998. Multivariate reduced-rank regression. Springer, New York.

Sarkis, M., Diepold, K., Westad, F., 2006. A new algorithm for gene mapping: Application of partial least squares regression with cross model validation. Genomic Signal Processing and Statistics, 2006. GENSIPS'06. IEEE International Workshop on, pp. 89–90.

Stein, J.L., Hua, X., Lee, S., Ho, A.J., Leow, A.D., Toga, A.W., Saykin, A.J., Shen, L., Foroud, T., Pankratz, N., Huentelman, M.J., Craig, D.W., Gerber, J.D., Allen, A.N., Corneveaux, J.J., Dechairo, B.M., Potkin, S.G., Weiner, M.W., Thompson, P., 2010. Voxelwise genome-wide association study (vGWAS). NeuroImage. .

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58 (1), 267–288.

Tziortzi, A., Searle, G., Tzimopoulou, S., Salinas, C., Beaver, J., Jenkinson, M., Rabiner, E., and Gunn, R. (2010). Imaging dopamine receptors in humans with [11C]-(+)-phno: dissection of d3 signal and anatomy. NeuroImage (in submission).

Waaijenborg, S., de Witt Hamer, V., Philip, C., Zwinderman, A., 2008. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. Statistical Applications in Genetics and Molecular Biology 7 (1), 3.

Wang, K., Abbott, D., 2008. A principal components regression approach to multilocus genetic association studies. Genetic Epidemiology 32 (2), 108–118.

Wang, W., Barratt, B., Clayton, D., Todd, J., 2005. Genome-wide association studies: theoretical and practical concerns. Nature Reviews Genetics 6 (2), 109–118.

Witten, D., Tibshirani, R., Hastie, T., 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10 (3), 515.

Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E., Lange, K., 2009. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics (Oxford, England) 25 (6), 714–721.

Please cite this article as: Vounou, M., et al., Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.07.002