The Canadian Journal of Statistics
Vol. 47, No. 1, 2019, Pages 108–131
La revue canadienne de statistique

A review of statistical methods in imaging genetics

Farouk S. NATHOO[1*], Linglong KONG[2], and Hongtu ZHU, for the Alzheimer's Disease Neuroimaging Initiative[3]

[1]*Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia, Canada*
[2]*Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada*
[3]*Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, U.S.A.*

Key words and phrases: Group sparse regression; multivariate high-dimensional regression; neuroimaging phenotype; single nucleotide polymorphism; reduced rank regression.

MSC 2010: Primary 62F15; secondary 62G10

Abstract: With the rapid growth of modern technology, many biomedical studies are being conducted to collect massive datasets with volumes of multi-modality imaging, genetic, neurocognitive and clinical information from increasingly large cohorts. Simultaneously extracting and integrating rich and diverse heterogeneous information in neuroimaging and/or genomics from these big datasets could transform our understanding of how genetic variants impact brain structure and function, cognitive function and brain-related disease risk across the lifespan. Such understanding is critical for diagnosis, prevention and treatment of numerous complex brain-related disorders (e.g., schizophrenia and Alzheimer's disease). However, the development of analytical methods for the joint analysis of both high-dimensional imaging phenotypes and high-dimensional genetic data, a big data squared (BD^2) problem, presents major computational and theoretical challenges for existing analytical methods. Besides the high-dimensional nature of BD^2, various neuroimaging measures often exhibit strong spatial smoothness and dependence and genetic markers may have a natural dependence structure arising from linkage disequilibrium. We review some recent developments of various statistical techniques for imaging genetics, including massive univariate and voxel-wise approaches, reduced rank regression, mixture models and group sparse multi-task regression. By doing so, we hope that this review may encourage others in the statistical community to enter into this new and exciting field of research. *The Canadian Journal of Statistics* 47: 108–131; 2019 © 2019 Statistical Society of Canada

Résumé: Avec l'évolution rapide de la technologie, de nombreuses études biomédicales collectent des jeux de données massifs comportant un volume d'images multi-modales et des informations cliniques, génétiques et neurocognitives sur des cohortes de plus en plus grandes. Réussir à en extraire puis à intégrer simultanément des informations riches et hétérogènes en génomique ou en imagerie cérébrale pourrait transformer notre compréhension des conséquences de la génétique sur les structures du cerveau et ses fonctions, cognitives ou autres, ainsi que sur les maladies cérébrales affectant les individus au cours de leur vie. Cette compréhension est cruciale pour le diagnostic, la prévention et le traitement de nombreux troubles cérébraux complexes (comme la schizophrénie et la maladie d'Alzheimer). L'analyse conjointe de phénotypes mesurés par l'imagerie en haute dimension avec des données génétiques également en haute dimension mène à un problème de mégadonnées au carré (MD^2), présentant des défis computationnels et théoriques. Au-delà de la haute dimension de données MD^2, les mesures d'imagerie médicale comportent souvent une dépendance spatiale et une apparence lisse, puis les marqueurs génétiques peuvent posséder

---

Additional Supporting Information may be found in the online version of this article at the publisher's website.

* *Corresponding author.*
E-mail: nathoo@uvic.ca

une structure de dépendance naturelle émergeant du déséquilibre des liens. Les auteurs décrivent le développement récent de plusieurs techniques statistiques pour l'imagerie en génétique, notamment les approches univariée massive et par voxel, la régression de rang réduit, les modèles de mélange, et la régression multi-tâches pour groupes épars. Ils souhaitent ainsi encourager d'autres membres de la communauté statistique à contribuer à cet excitant nouveau champ de recherche. *La revue canadienne de statistique* 47: 108–131; 2019 © 2019 Société statistique du Canada

1. INTRODUCTION

Despite the numerous successes of genome-wide association studies (GWAS), it has been difficult to unravel the genetic basis of many complex neurological diseases since each genetic variant may only contribute in a small way to disease risk and such a genetic basis can be very heterogeneous (Cannon & Keller, 2006; Peper et al., 2007; Marenco & Radulescu, 2010). The additive and interactive effects of perhaps hundreds of risk genes and multiple environmental risk factors, each with small individual effects, may contribute to the abnormal developmental trajectories that underlie neurological and psychiatric disorders such as Alzheimer's disease (AD). Identifying such risk genes and environmental risk factors could transform our understanding of the origins of these conditions and inspire new approaches for urgently needed preventions, diagnoses and treatments. Once such an identification has been accomplished, lifestyle and medical interventions can be applied.

A promising approach to understanding the genetic basis of neurological disorders is through studies that integrate multi-scale data from genetic/genomic, multimodal brain imaging and environmental risk factors (Hibar et al., 2011; Thompson et al., 2013; Hibar et al., 2015), so called imaging genetics studies.

To promote such studies, the Brain Imaging Clinical Research Program at the National Institute of Mental Health (NIMH) has called for the investigation of relationships between genetic variations and imaging and cognitive findings and phenotypes in adult mental disorders. To this end, a number of large-scale publicly available imaging genetic databases have been established, including the Human Connectome Project (HCP) study, the UK biobank (UKbb) study, the Pediatric Imaging, Neurocognition and Genetics (PING) study, the Philadelphia Neurodevelopmental Cohort (PNC) and the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, among many others. The ADNI database in particular has been used extensively by statisticians working on the development of methods for the joint analysis of neuroimaging and genetic data, and this database serves as a good starting point for new researchers in the area.

In these imaging genetic studies, the data available for each subject may include multiple magnetic resonance imaging (MRI) images, such as structural MRI, diffusion tensor imaging (DTI) and functional MRI (fMRI), as well as cognitive assessments, and genomic data [e.g., single nucleotide polymorphism (SNP) array and copy number variations (CNVs)]. Jointly analyzing neuroimaging and genetic data with clinical variables, however, raises serious challenges as existing statistical methods are rendered infeasible for efficiently analyzing large-scale imaging genetic data sets with many subjects. These challenges arise from a setting where the data involve high-dimensional imaging data $\times$ high-dimensional genetic data−so-called big data squared (BD^2), complex correlation and spatial structures within both imaging and genetic data and a potentially large number of subjects.

For many brain-related diseases, since changes in brain structure and function are very subtle, it is common to normalize multi-modal neuroimaging data to a common template (Miller & Younes, 2001; Xu et al., 2003). After normalization, various imaging phenotypes (IPs) are commonly calculated from structural and functional imaging data (Zhu, Gu, & Peterson, 2003; Friston, 2009). These normalized neuroimaging phenotypes are functional data measured at a very large number ($10^4 - 10^7$) of grid points along space and/or time and network data measured among a large number ($10^4 - 10^6$) of region of interest (ROI) pairs (Hibar et al., 2011; Thompson

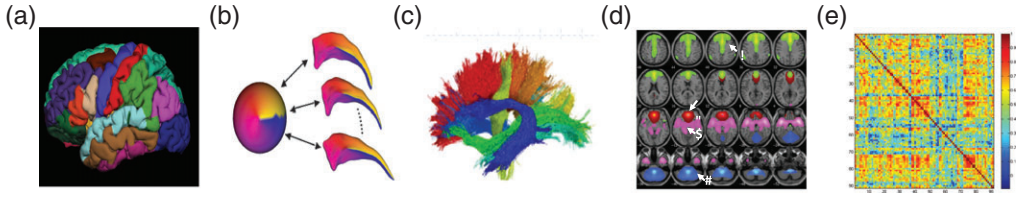

FIGURE 1: Selected Imaging Phenotypes: (a) a cortical parcellation; (b) shape features; (c) features derived from diffusion tensor imaging; (d) MRI; and (e) correlation matrix obtained from functional neuroimaging data.

et al., 2013; Hibar et al., 2015; Ge et al., 2015). See Figure 1 for a graphical depiction of potential IPs.

The earliest methods developed for imaging genetics data analysis are based on significant reductions to both data types, for example, restricting the analysis to a specific candidate ROI in the brain and/or a specific candidate genetic marker. This type of univariate analysis can be extended to handle full brain-wide genome-wide data based on the application of a massive number of pairwise univariate analyses, each based on a standard linear regression relating a given voxel or region to a given SNP. In this case a great deal of hypotheses are tested at once, and the resulting multiple testing corrections are very strict. For example, in a full brain-wide genome-wide study involving 10^6 known genetic variants and 10^6 locations in the brain, this type of analysis will require 10^{12} univariate analyses. Furthermore, the resulting *P*-values are not independent because of spatial correlation in the imaging data.

Stein et al. (2010) are the first to consider such an analysis. These authors examine 448,293 SNPs in each of 31,622 voxels of the entire brain across 740 elderly subjects. A total of 300 computational-cluster nodes are used to carry out the required computations in parallel. Hibar et al. (2011) consider a similar analysis but reduce the number of tests by conducting the analysis at the gene rather than SNP level. In this case principal component analysis (PCA) is used to summarize the SNP data for each gene, and the resulting ''eigenSNPs'' are used in the massive univariate analysis.

As an alternative to the massive univariate approach, a voxel-wise approach continues to fit regression models separately at each location in the brain, but considers a set of genetic markers simultaneously rather than just a single genetic marker. Ge et al. 2012 develop such an analysis and examine a dataset that is similar to that considered in the article by Stein et al. (2010), but a key difference is the use of a multi-locus model based on least squares kernel machines (Liu, Lin, & Ghosh, 2007), which is used to combine the effect of multiple genetic variants and model their interaction. In addition, the spatial information in the images is accounted for through the use of random field theory (RFT) as an inferential tool (Worsley, 2002). This approach is extended in the article by Ge et al. (2015) to allow for potential interactions between genetic variables and non-genetic variables such as disease-risk factors, environmental exposures and epigenetic markers.

An alternative fast voxel-wise genome-wide association analysis (FVGWAS) approach is developed by Huang et al. (2015) where the authors focus on reducing the computational burden required for a full brain-wide gene-wide study. This objective is implemented in part by incorporating a global sure independence screening procedure along with inference based on the wild bootstrap. The resulting approach can implement a brain-wide genome-wide analysis in a relatively small amount of time utilizing only a single CPU.

One drawback of the massive univariate and voxel-wise approaches is that the relationship between the different neuroimaging phenotypes (e.g., at different regions of the brain) is not

explicitly modelled, and therefore, potential efficiency gains arising from the borrowing of information across brain regions are not realized. An alternative approach is to base the analysis on a single large model, a multivariate high-dimensional regression model that is fit to the entire dataset. In this framework the scale of the data must necessarily be reduced, and it is common to summarize the entire image using a relatively moderate number of brain summary measures across some key ROIs. As an example, Table 1 describes a phenotype of dimension 56 that can be derived from an MRI image, and these data are considered in our example application.

One such regression model is the group-sparse multitask regression model proposed by Wang et al. (2012a, 2012b) where estimation of the regression coefficients in a multivariate linear model is based on penalized least squares. The penalty is chosen to induce a particular form of structured sparsity in the solutions based on two nested forms of grouping. The first is at the SNP level (grouping the regression coefficients of a given SNP across all phenotypes) and the second is at the gene level, which groups all SNPs within a given gene. More recently, Greenlaw et al. (2017) have extended this approach to the Bayesian setting which allows for inference and uncertainty quantification for the regression coefficients of the selected genetic markers.

An alternative multivariate approach is based on approximating the high-dimensional regression coefficient matrix with a low-rank matrix. Such an approach has been developed by Vounou et al. (2010), who construct a sparse reduced-rank regression (SRRR) model which is applied to an imaging genetics study involving 111 anatomical ROIs and 437,577 SNPs. Using simulation studies, Vounou et al. (2010) show that their SRRR model has higher power to detect important genetic variants compared with the massive univariate approach. Along similar lines, Zhu et al. (2014) also develop a low rank regression model with inference conducted in the Bayesian framework and they apply their approach to an imaging genetics study involving 93 ROIs and 1,071 SNPs. Also in the Bayesian framework, Stingo et al. (2013) develop a hierarchical mixture model for relating brain connectivity to genetic information for studies involving fMRI data. The mixture components of the proposed model are used to classify the study subjects into subgroups, and the allocation of subjects to these mixture components is linked to genetic markers with regression parameters assigned spike-and-slab priors. The proposed model is used to examine the relationship between functional brain connectivity based on fMRI data and genetic variation. We note that the word functional is used here to refer to brain function and not to functional data.

Huang et al. (2017) develop a functional genome-wide association analysis (FGWAS) framework to efficiently carry out whole-genome analyses of phenotypes measuring brain function. Compared with FVGWAS, FGWAS explicitly models the features of these phenotypes through the integration of smooth coefficient functions and functional PCA. In the latter context the word functional refers to functional data and not brain function. Statistically, compared with existing methods for GWAS, FGWAS can substantially boost the detection power for discovering important genetic variants influencing brain structure and function.

In more recent work, researchers have turned their attention to longitudinal imaging genetics studies where study subjects are followed over time with neuroimaging data collected over a sequence of time points during a follow-up period. With longitudinal MRI data, changes in the structure of the brain over time can be characterized, for example, by examining rates of brain deterioration, and these estimated rates of change can be related to genetic markers. Szefer et al. (2017) examine the presence of linear association between minor allele counts of 75,845 SNPs in the Alzgene linkage regions and estimated rates of change of structural MRI measurements for 56 brain regions. The authors develop a bootstrap-enhanced sparse canonical correlation analysis (SCCA) to create refined lists of SNPs associated with rates of structural change over time.

Lu et al. (2017) develop a Bayesian approach to perform longitudinal analysis of multivariate neuroimaging phenotypes and candidate genetic markers obtained from longitudinal studies. A low rank longitudinal regression model is specified where penalized splines are incorporated

TABLE 1: Imaging phenotypes defined as volumetric or cortical thickness measures of $28 \times 2 = 56$ ROIs from automated Freesurfer parcellations. Each of the phenotypes in the table corresponds to two phenotypes in the data: one for the left hemisphere and the other for the right hemisphere.

ID	Measurement	Region of interest
AmygVol	Volume	Amygdala
CerebCtx	Volume	Cerebral cortex
CerebWM	Volume	Cerebral white matter
HippVol	Volume	Hippocampus
InfLatVent	Volume	Inferior lateral ventricle
LatVent	Volume	Lateral ventricle
EntCtx	Thickness	Entorhinal cortex
Fusiform	Thickness	Fusiform gyrus
InfParietal	Thickness	Inferior parietal gyrus
InfTemporal	Thickness	Inferior temporal gyrus
MidTemporal	Thickness	Middle temporal gyrus
Parahipp	Thickness	Parahippocampal gyrus
PostCing	Thickness	Posterior cingulate
Postcentral	Thickness	Postcentral gyrus
Precentral	Thickness	Precentral gyurs
Precuneus	Thickness	Precuneus
SupFrontal	Thickness	Superior frontal gyrus
SupParietal	Thickness	Superior parietal gyrus
SupTemporal	Thickness	Superior temporal gyrus
Supramarg	Thickness	Supramarginal gyrus
TemporalPole	Thickness	Temporal pole
MeanCing	Mean thickness	Caudal anterior cingulate, isthmus cingulate, posterior cingulate, rostral anterior cingulate
MeanFront	Mean thickness	Caudal midfrontal rostral midfrontal, superior frontal lateral orbitofrontal, and medial orbitofrontal gyri frontal pole
MeanLatTemp	Mean thickness	Inferior temporal, middle temporal superior temporal gyri
MeanMedTemp	Mean thickness	Fusiform, parahippocampal, and lingual gyri, temporal pole and transverse temporal pole
MeanPar	Mean thickness	Inferior and superior parietal gyri supramarginal gyrus, and precuneus
MeanSensMotor	Mean thickness	Precentral and postcentral gyri
MeanTemp	Mean thickness	Inferior temporal, middle temporal, superior temporal, fusiform, parahippocampal, lingual gyri, temporal pole, transverse temporal pole

to characterize an overall time effect, and a sparse factor analysis model coupled with random effects is proposed to account for spatiotemporal correlations of longitudinal phenotypes. A useful feature of the proposed methodology is the allowance for interactions between genetic main effects and time.

In the remainder of the article, Sections 2–4 discuss in more detail some of the methods mentioned above. Section 5 presents an example application where data from the ADNI-1 database are used to examine the association between the 56 neuroimaging phenotypes presented in Table 1 and a collection of 486 SNPs from 33 genes belonging to the top 40 AD candidate genes listed on the AlzGene database as of June 10, 2010. Section 6 concludes with a discussion of some ongoing work in this area.

2. MASS UNIVARIATE AND VOXEL-WISE APPROACHES

Mass univariate approaches avoid the complication of jointly modelling all neuroimaging phenotypes and genetic markers and simply conduct a test for association at each possible pair of voxel and genetic marker. Voxel-wise approaches are similar in that a separate model is fit independently at each voxel of the image, but these approaches may include multiple genetic markers in each model. The primary advantage of these approaches is that they make feasible a full brain-wide and genome-wide search for associations without the need to reduce the imaging data to a smaller number of ROIs.

We assume that neuroimaging and genetic data are available on n subjects, where the IP is denoted as $y_\ell(v)$, for the numerical value of the brain image of subject ℓ, $\ell = 1, \ldots, n$, at voxel $v, v = 1, \ldots, V$. We denote the set of genetic markers for subject l by $\mathbf{x}_\ell = (x_{\ell 1}, \ldots, x_{\ell d})^T$, $\ell = 1, \ldots, n$, for a total of d markers, where $x_{\ell j} \in \{0, 1, 2\}$ is the number of copies of the minor allele for the jth SNP, which takes values $x_{\ell j} = 0$ (homozygotic major alleles), $x_{\ell j} = 1$ (heterozygote) and $x_{\ell j} = 2$ (homozygotic minor alleles). Finally, we let $\mathbf{z}_\ell = (z_{\ell 1}, \ldots, z_{\ell p})^T$, $\ell = 1, \ldots, n$, denote a collection of non-genetic variables for subject l.

Stein et al. (2010) is the first voxel-wise genome-wide association study (vGWAS) examining genetic influence on brain structure. The authors consider neuroimaging and genetic data obtained from $n = 818$ subjects as part of the ADNI. The neuroimaging data are based on brain MRI scans that are processed using an approach known as tensor-based morphometry (TBM). TBM (Ashburner, Good, & Friston, 2000) is applied to each of the MRI scans to create images representing volumetric tissue differences at each of approximately 31,000 voxels for each individual, where the value of the image in a given voxel is obtained by calculating the determinant of the Jacobian matrix of a deformation that encodes local volume excess or deficit relative to an image that is representative of the sample, known as the mean template image. The analysis relates the value of the image at each voxel to each of 448,293 SNPs.

The statistical methodology considered by Stein et al. (2010) is fairly straightforward, though the resulting computation is still extensive due to the total number of tests considered. At each voxel v, a linear regression is conducted with response $y_\ell(v)$ (volumetric tissue difference relative to a mean template image at voxel v) and a separate regression model is fit relating this response to each SNP $x_{\ell j}$, assuming an additive genetic model. Additional independent variables age and gender are also included in the model,

$$y_\ell(v) = \beta_0 + \beta_1 \text{Age}_\ell + \beta_2 \text{Gender}_\ell + \alpha x_{\ell j} + e_\ell(v, j).$$

A standard P-value from this linear regression is obtained for each SNP-voxel pair (corresponding to a null hypothesis of $\alpha = 0$), and these P-values are computed at a given voxel as the model is fit repeatedly to all d SNPs.

To conserve memory, Stein et al. (2010) only save the minimum P-value at each voxel. Under the null hypothesis that the phenotype at a given voxel is not associated with any of the

genetic markers, the minimum P-value computed at each voxel is not uniform$[0, 1]$, but it is shown to be approximately Beta$(1, M_{eff})$, with $M_{eff} < M$, where M_{eff} is the effective number of independent tests conducted at each voxel, and M is the total number of genetic markers. The inequality $M_{eff} < M$ arises as a result of linkage disequilibrium.

Stein et al. (2010) set the value of M_{eff} equal to the number of principal components required to jointly explain 99.5% of the variance in the SNPs. The Beta$(1, M_{eff})$ distribution is used to correct the minimum P-value computed at each voxel via the probability integral transform so that the corrected minimum P-value is approximately distributed as uniform$[0, 1]$. False discovery rate (FDR) procedures (Benjamini & Hochberg, 1995) are then applied to adjust for multiple testing across voxels. Under the proposed scheme the computations can be carried out in parallel across voxels, and Stein et al. (2010) employ 300 cluster nodes with a reported 27 h of total computation time.

Hibar et al. (2011) develop a gene-based association method to complement single-marker GWAS for implicating underlying genetic variants in complex traits and diseases. The authors focus more broadly on gene-based approaches as they can be more powerful than traditional SNP-based approaches, with the relative power depending on how the genetic variants affect the phenotype. For example, if a gene contains multiple causal variants with small individual effects, SNP-based methods may miss these associations if a very stringent significance threshold is used. Gene-based tests also reduce the effective number of statistical tests by aggregating multiple SNP effects into a single test statistic.

In Hibar et al. (2011) the SNP data are grouped by gene and SNPs that are not located in a gene are excluded. Considering a dataset of the same scale, both in terms of imaging and genetics, as that considered in Stein et al. (2010), after grouping SNPs a total of 18,044 genes are left for analysis. The authors propose a gene-based association method that is based on principal components regression. PCA is performed on the SNP data within each gene, storing all of the orthonormal basis vectors of the SNP matrix that explain a majority of the variance in the set of SNPs for a given gene. Basis vectors with the highest eigenvalues (higher proportions of explained variance) are included until 95% of the SNP variability within the gene is explained, and the rest are discarded. The resulting ''eigenSNPs'' approximate the information in the collection of observed SNPs for a given gene.

Hibar et al. (2011) apply their approach by examining associations between 18,044 genes and approximately 31,000 voxel-specific phenotypes. At each voxel, a multiple partial-F test is employed to test the joint effect of all eigenSNPs for a given gene on the value of the image (volume difference) at the given voxel. The test is applied to all genes and the minimum P-value is recorded at each voxel. Inference then proceeds using an approximate Beta null distribution with FDR procedures applied to adjust for multiple testing as in Stein et al. (2010). The required computation across all voxels is parallelized over a cluster of 10 high performance 8-core CPU nodes, and Hibar et al. (2011) report that the total time required to complete an analysis with their computational setup is approximately 13 days. Summarizing the SNP information in this way may have some disadvantages as well. In particular, if a single SNP has a large main effect, then testing the joint effect of all SNPs within that gene may dilute this association. However, when one considers the drastic reduction in the number of independent tests when comparing SNP-based linear regression with gene-specific summaries based on PCA, gene-based testing offers advantages when dealing with an extremely large number of voxels in an image phenotype.

In more recent work, Huang et al. (2015) have developed a FVGWAS analysis with an emphasis on large-scale imaging and genetic data. A key advantage of this methodology over the methods developed by Stein et al. (2010) and Hibar et al. (2011) is that it requires considerably less computational resources and is feasible to run on just a single CPU with reasonable processing time. The proposed approach is based on three main components: (i) a heteroscedastic

linear model is specified at each voxel-locus pair; (ii) a global sure independence screening procedure is incorporated to eliminate many noisy loci; and (iii) inference is based on wild bootstrap methods. The heteroscedastic linear model at voxel v and locus c takes the form

$$y_\ell(v) = \mathbf{z}_\ell^T \boldsymbol{\beta}(v) + x_{\ell c}\alpha(c, v) + e_\ell(v), \ \ell = 1, \ldots, n$$

and the model makes no strong assumptions on $\mathrm{Var}[e_\ell(v)]$, in particular, it may vary across subjects. The hypothesis test of interest is

$$H_0(c, v) : \alpha(c, v) = 0 \text{ versus } H_1(c, v) : \alpha(c, v) \neq 0 \text{ for each } (c, v).$$

Huang et al. (2015) introduce a standard Wald statistic $W(c, v)$ that is based on the ordinary least squares estimate of $\alpha(c, v)$. Under the heteroscedastic assumption of the regression model the standard approximations to the null distribution of $W(c, v)$ based on the χ_1^2 (or F) distribution do not apply and a wild bootstrap method is proposed as an alternative. Huang et al. (2015) then focus on approximations that make such a procedure computationally feasible.

A key aspect of these approximations is that a global sure independence screening procedure is used to eliminate many noisy loci. The global aspect of the screening procedure reduces the set of SNPs *for all voxels simultaneously*. The authors define a global Wald statistic $W(c)$ for a given locus as the average of $W(c, v)$ taken over all voxels in the image. If, for a given locus c, it is the case that $H_0(c, v)$ holds for all voxels v, then Huang et al. (2015) argue that $W(c, v)$ asymptotically converges to a weighted χ^2 distribution. The corresponding P-values are then computed for each locus c, and the top N_0 loci (e.g., $N_0 = 1,000$) are selected as the candidate set.

Given the candidate set, a wild bootstrap approach is applied to determine significant voxel-locus pairs, or alternatively, significant cluster-locus pairs, where a cluster refers to a set of interconnected voxels each with test statistic exceeding a certain threshold. Allowing for cluster-wise inference in this way is an important advantage of this methodology over that proposed by Stein et al. (2010) and Hibar et al. (2011), as the voxel-specific tests proposed in the latter two articles might miss spatially extended signals that do not achieve significance at any isolated voxel. In this sense Huang et al. (2015) take advantage of the spatial information in the 3D images.

Ge et al. (2012) develop the first voxel-wise imaging genetics approach that allows for interactions between genetic markers. At each voxel the authors propose to fit a multi-locus model to associate the joint effect of several SNPs with the imaging trait at that voxel. The imaging traits are similar to those considered in Stein et al. (2010) though the model specified at each voxel is different. In particular, the semiparametric regression model specified at each voxel takes the form

$$y_\ell(v) = \mathbf{z}_\ell^T \boldsymbol{\beta}(v) + h_v(\mathbf{x}_\ell) + e_\ell(v), \ \ell = 1, \ldots, n,$$

where $h_v(\mathbf{x}_\ell)$ denotes a nonparametric function of the SNPs and the errors are assumed to be normally distributed with mean 0 and standard deviation σ_v. In this case the non-genetic covariates [e.g., age, gender, education, handedness and total intracranial volume (ICV)] are modelled parametrically and the effect of genetic markers is modelled nonparametrically using a least squares kernel machines (Liu, Lin, & Ghosh, 2007) approach. The function space containing $h_v(\cdot)$ is determined by an $n \times n$ kernel matrix which is a function of the genetic data and must be positive definite. The (j, k) element of this matrix is a function of the SNP genotypes of subjects j and k, and Ge et al. (2012) specify the form of this kernel to be

$$k(\mathbf{x}_j, \mathbf{x}_k) = \frac{1}{2d} \sum_{s=1}^{d} \mathrm{IBS}(x_{js}, x_{ks})$$

where $\text{IBS}(x_{js}, x_{ks})$ denotes the number of alleles shared identical by descent by subjects j and k at SNP s and takes values 0, 1 or 2.

In this case the null hypothesis of interest is $H_0(v) : h_v(\cdot) = 0$, which examines the effect of multiple SNPs at each voxel. Importantly, the model is very flexible and allows for interactions between the genetic markers. Ge et al. (2012) exploit a connection between least squares kernel machines and linear mixed models to derive a score statistic based on the null model, that is, the model with no effects from SNPs, and argue that this statistic follows a mixture of chi-square distributions under the null hypothesis. The score statistic has the advantage that its computation does not require the estimation of the function $h(\cdot)$. Using the Satterthwaite method, the distribution of this statistic under the null hypothesis is approximated by a scaled chi-squared distribution.

Applying this technique at all voxels produces an image of score statistics and the authors assume that this statistic image behaves like a χ^2 random field which facilitates inference using RFT (Worsley et al., 1996). RFT produces FWE-corrected P-values for voxel-wise and cluster-wise inference by accounting for the volume and smoothness of the statistic image. As RFT requires fairly strong assumptions on the statistic image and these assumptions may not be satisfied, the authors also develop voxel-wise inference based on permutation procedures with a parametric tail approximation based on the generalized Pareto distribution.

Along with allowing for interactions among genetic variables the work of Ge et al. (2012) is the first to use RFT for inference in imaging genetics. Thus while correlation across voxels is not accounted for directly within the statistical model, the spatial structure of the imaging data is accounted for when computing FWE-corrected P-values using RFT.

In a subsequent article, Ge et al. (2015) extend the least squares kernel machine approach of Ge et al. (2012) to allow interactions between SNPs and further allow interactions between SNPs and non-genetic variables such as disease risk factors, environmental exposures, and epigenetic markers. The model specified is of the form

$$y_\ell(v) = \mathbf{z}_\ell^T \boldsymbol{\beta}(v) + h_{v,x}(\mathbf{x}_\ell) + h_{v,w}(\mathbf{w}_\ell) + h_{v,x,w}(\mathbf{x}_\ell, \mathbf{w}_\ell) + e_\ell(v), \;\; \ell = 1, \ldots, n,$$

where $\mathbf{z}_\ell$ represents non-genetic variables with linear effect and $\mathbf{w}_\ell$ represents non-genetic variables with nonlinear effect that may interact with the genetic markers. As before, a kernel machine-based method is used to represent the nonparametric effects. In their application, Ge et al. (2015) only consider a scalar phenotype derived through MRI, namely, the hippocampal volume averaged between the two brain hemispheres; however, combining the voxel-wise inference of Ge et al. (2012) with the more flexible kernel machine model of Ge et al. (2015) seems feasible for dealing with phenotypes comprising an entire 3D image.

The mass univariate and voxel-wise approaches are appealing because of their simplicity and because the required univariate or multi-locus regression models are relatively straightforward to fit. Modelling the dependence between different voxels is avoided and this makes it feasible to perform large scale searches across many voxels of an image. Despite these advantages an important limitation is that these approaches do not exploit the spatial structure of phenotype-genotype associations. If a particular SNP is related to one voxel then it will likely be related to the neighbouring voxels as well, and these approaches do not allow us to borrow information across voxels. This borrowing of information can lead to higher power and is thus desired. Multivariate approaches are thus natural to consider, but these models typically require a substantial reduction in the dimension of the neuroimaging phenotype by two orders of magnitude.

3. MULTIVARIATE APPROACHES

With multivariate approaches all of the neuroimaging phenotypes are included in a single large model that may account for the dependence structure across the different phenotypes while

relating each of the phenotypes to all of the genetic markers. As a result, these approaches are typically not applied to imaging data at the voxel level as this is computationally intractable. A multivariate approach is typically applied to images reduced to a much coarser scale where each phenotype corresponds to a summary measure for an ROI in the brain. Table 1 provides an example of such summary measures for the 56 ROIs considered in our example.

In the work of Wang et al. (2012a, 2012b) an estimator based on group sparse regularization is applied to multivariate regression for relating neuroimaging phenotypes to SNPs, where the SNPs are grouped by genes and this grouping structure is accounted for in the construction of the regression estimator. Let $\mathbf{y}_\ell = (y_{\ell 1}, \ldots, y_{\ell c})^T$ denote the IP summarizing the structure of the brain over c ROIs for subject ℓ, $\ell = 1, \ldots, n$. The corresponding genetic data are denoted by $\mathbf{x}_\ell = (x_{\ell 1}, \ldots, x_{\ell d})^T$, $\ell = 1, \ldots, n$, where we have information on d SNPs, and $x_{\ell j} \in \{0, 1, 2\}$ is the number of minor alleles for the jth SNP. We further assume that each SNP can be associated with a gene so that the set of genes represents a higher level grouping of the SNPs. Thus the set of SNPs can be partitioned into K genes, and we let $\pi_k, k = 1, 2, \ldots, K$, denote the set containing the SNP indices corresponding to the kth group and $m_k = |\pi_k|$. This partitioning is used to allow for gene-wise association among SNPs. This is done through a regularization in which the coefficients of the SNPs within a gene, with respect to all of the IPs, are penalized as a whole with an l_2-norm, while the l_1-norm is used to sum up the gene-wise penalties to enforce sparsity between genes. The latter is important because in reality only a small fraction of genotypes are related to a specific phenotype.

It is assumed that $E(\mathbf{y}_\ell) = \mathbf{W}^T \mathbf{x}_\ell$, $\ell = 1, \ldots, n$, where $\mathbf{W}$ is a $d \times c$ matrix, with each row characterizing the association between a given SNP and the brain summary measures across all c ROIs. The estimator proposed by Wang et al. (2012a, 2012b) takes the form

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_{\ell=1}^{n} ||\mathbf{W}^T \mathbf{x}_\ell - \mathbf{y}_\ell||_2^2 + \gamma_1 ||\mathbf{W}||_{G_{2,1}} + \gamma_2 ||\mathbf{W}||_{l_{2,1}}, \tag{1}$$

where γ_1 and γ_2 are regularization parameters weighting a $G_{2,1}$-norm penalty $||\mathbf{W}||_{G_{2,1}} = \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2}$ and an $\ell_{2,1}$-norm penalty $||\mathbf{W}||_{l_{2,1}} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} w_{ij}^2}$ respectively. The $G_{2,1}$-norm encourages sparsity at the gene level. This regularization differs from group LASSO (Yuan & Lin, 2003) as it penalizes regression coefficients for a group of SNPs across all IPs jointly. As an important gene may contain irrelevant individual SNPs, or a less important gene may contain individually significant SNPs, the second penalty, an $\ell_{2,1}$-norm (Argyriou, Evgeniou & Pontil, 2007), is added to allow for additional structured sparsity at the level of SNPs (the rows of $\mathbf{W}$).

The estimator (1) provides a novel approach for assessing associations between neuroimaging phenotypes and genetic variations as it accounts for several interrelated structures within genotyping and imaging data. Wang et al. (2012a, 2012b) develop an optimization algorithm for the computation of (1) and suggest the use of cross-validation for the selection of tuning parameters γ_1 and γ_2. A limitation of the proposed methodology is that it only furnishes a point estimate $\hat{\mathbf{W}}$ and techniques for obtaining valid standard errors or interval estimates are not provided.

In recent work, Greenlaw et al. (2017) address this limitation and extend the methodology of Wang et al. (2012a, 2012b) so that the uncertainty associated with $\hat{\mathbf{W}}$ can be quantified. Their methodology allows for formal statistical inference beyond the sparse point estimate $\hat{\mathbf{W}}$. Following the ideas of Park & Casella (2008) and Kyung et al. (2010), Greenlaw et al. (2017) develop a hierarchical Bayesian model that allows for full posterior inference. The Bayesian model is constructed with a particular prior for $\mathbf{W}$ so that the estimator (1) corresponds to the posterior mode. The spread of the posterior distribution then provides valid measures

of posterior variability along with credible intervals and/or posterior probabilities for each regression parameter.

Let $\mathbf{W}^{(k)} = (w_{ij})_{i \in \pi_k}$ denote the $m_k \times c$ submatrix of $\mathbf{W}$ containing the rows corresponding to the kth gene, $k = 1, \ldots, K$. The hierarchical model of Greenlaw et al. (2017) corresponding to the estimator (1) takes the form

$$\mathbf{y}_\ell \,|\mathbf{W}, \sigma^2 \overset{ind}{\sim} MVN_c(\mathbf{W}^T \boldsymbol{x}_\ell \,,\, \sigma^2 I_c), \quad \ell = 1, \ldots, n, \tag{2}$$

with the coefficients corresponding to different genes assumed to be conditionally independent

$$\mathbf{W}^{(k)} | \lambda_1^2, \lambda_2^2, \sigma^2 \overset{ind}{\sim} p(\mathbf{W}^{(k)} | \lambda_1^2, \lambda_2^2, \sigma^2) \quad k = 1, \ldots, K, \tag{3}$$

and with the prior distribution for each $\mathbf{W}^{(k)}$ having a density function given by

$$\begin{aligned} p(\mathbf{W}^{(k)} | \lambda_1^2, \lambda_2^2, \sigma^2) \propto \exp\left\{ -\frac{\lambda_1}{\sigma} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2} \right\} \\ \times \prod_{i \in \pi_k} \exp\left\{ -\frac{\lambda_2}{\sigma} \sqrt{\sum_{j=1}^{c} w_{ij}^2} \right\}. \end{aligned} \tag{4}$$

By construction, the posterior mode, conditional on $\lambda_1^2, \lambda_2^2, \sigma^2$, corresponding to the model hierarchy (2)–(4) is exactly the estimator (1) proposed by Wang et al. (2012a, 2012b) with $\gamma_1 = 2\sigma\lambda_1$ and $\gamma_2 = 2\sigma\lambda_2$. This equivalence between the posterior mode and the estimator of Wang et al. (2012a, 2012b) is the motivation for the model; however, generalizations that allow for a more flexible covariance structure in (2) can also be considered, and an extension of this model to allow for spatial correlation is discussed in Section 6.

Greenlaw et al. (2017) develop a Gaussian scale mixture representation of this hierarchical model which allows for the implementation of Bayesian inference using a straightforward Gibbs sampling algorithm that is implemented in the R package ‘‘bgsmtr’’ (Bayesian Group Sparse Multi-Task Regression) which is available for download on CRAN (https://cran.r-project.org/web/packages/bgsmtr/). The selection of the tuning parameters λ_1^2 and λ_2^2 for this model is investigated in Nathoo, Greenlaw, & Lesperance (2016), where selection of these tuning parameters based on a fully Bayes approach with hyperpriors, an empirical Bayes approach, and the WAIC are compared.

Vounou et al. (2010) propose an alternative strategy for multivariate regression modelling with imaging genetics data where the high-dimensional regression coefficient matrix is approximated by a low rank sparse matrix leading to a SRRR model. Suppose that $\mathbf{X}$ is the $n \times d$ design matrix of genetic markers and $\mathbf{Y}$ is the $n \times c$ matrix of IPs. Beginning with the standard multivariate multiple linear regression model $\mathbf{Y} = \mathbf{XC} + \mathbf{E}$, where $\mathbf{C}$ is the $d \times c$ matrix of regression coefficients, the approach proceeds by first imposing a rank condition on this matrix, $\text{rank}(\mathbf{C}) \leq \min(d, c)$, which leads to a decrease in the number of parameters that need to be estimated. In particular, if $\mathbf{C}$ has rank r then it can be expressed as $\mathbf{C} = \mathbf{BA}$ where $\mathbf{B}$ is $d \times r$ and $\mathbf{A}$ is $r \times c$ such that $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{B}) = r$. The loss function for estimation is based on the weighted least squares criterion $M = \text{Tr}\{(\mathbf{Y} - \mathbf{XBA})\Gamma(\mathbf{Y} - \mathbf{XBA})^T\}$, where Γ is a $c \times c$ positive definite weight matrix. Vounou et al. (2010) consider sparse estimation of both $\mathbf{B}$ and $\mathbf{A}$ through penalized estimation incorporating l_1-norm penalties. In particular, setting Γ to be the identity matrix we have

$$M = \text{Tr}\{\mathbf{YY}^T\} - 2\text{Tr}\{\mathbf{AY}^T\mathbf{XB}\} + \text{Tr}\{\mathbf{AA}^T\mathbf{B}^T\mathbf{B}\},$$

where the first term on the RHS can be ignored as it does not depend on $\mathbf{B}$ or $\mathbf{A}$. Assuming $r = 1$ and adding l_1-norm penalization yields the following optimization problem.

$$\arg\min_{\mathbf{a},\mathbf{b}}\{-2\mathbf{a}\mathbf{Y}^T\mathbf{X}\mathbf{b} + \mathbf{a}\mathbf{a}^T\mathbf{b}^T\mathbf{b} + \lambda_a||\mathbf{a}^T||_1 + \lambda_b||\mathbf{b}^T||_1\}.$$

where $\mathbf{a}$ is $1 \times c$ corresponding to the phenotypes and $\mathbf{b}$ is $d \times 1$ corresponding to the genetic markers. The sparsity of the solution depends on the values of λ_a and λ_b with the non-zero elements of $\hat{\mathbf{a}}$ selecting phenotypes and the non-zero elements of $\hat{\mathbf{b}}$ selecting genetic markers.

The optimization problem is biconvex and Vounou et al. (2010) present an iterative algorithm for solving it. After the rank-one solution has been found, additional ranks can be obtained by applying the algorithm to the residuals of the data matrices. Vounou et al. (2010) suggest a graphical approach based on the residuals at each successive rank that can be used to select an optimal rank. As with the Wang et al. (2012a, 2012b) methodology, the methodology of Vounou et al. (2010) provides selection and point estimation but does not provide any mechanism for uncertainty quantification. This lack of uncertainty quantification can be a serious problem as we illustrate in our example application of Section 5.

Zhu et al. (2014) develop a Bayesian reduced rank model for imaging genetics that incorporates several enhancements above and beyond the methodology proposed in Vounou et al. (2010). First, the Bayesian approach enables uncertainty quantification for the regression parameters based on the posterior distribution. Second, in addition to reducing the dimension of the regression coefficient matrix with a low rank approximation, Zhu et al. (2014) also incorporate a sparse latent factor model to represent the high-dimensional covariance matrix of the brain IPs, with a multiplicative gamma process shrinkage prior assigned to the factor loadings.

As with Vounou et al. (2010) the proposed model is based on the multivariate linear model $\mathbf{Y} = \mathbf{X}\mathbf{C} + \mathbf{E}$, where $\mathbf{E} = (\epsilon_{\ell k})$ and the rows of $\mathbf{E}$, each corresponding to a different subject, are assumed independent with $\boldsymbol{\epsilon}_\ell \sim MVN_c(\mathbf{0}, \Sigma)$. The rank r decomposition of $\mathbf{C}$ with $r \ll \min(c, d)$ takes the form $\mathbf{C} = \sum_{j=1}^r \mathbf{C}_j = \sum_{j=1}^r \delta_j \mathbf{u}_j \mathbf{v}_j^T$, where $\mathbf{C}_j = \delta_j \mathbf{u}_j \mathbf{v}_j^T$ is the jth layer, $\mathbf{u}_j \in \mathbb{R}^d$ and $\mathbf{v}_j \in \mathbb{R}^c$. The regression errors for each subject are expressed using a latent factor model $\boldsymbol{\epsilon}_\ell = \boldsymbol{\Lambda}\boldsymbol{\eta}_\ell + \boldsymbol{\xi}_\ell$, where $\boldsymbol{\Lambda}$ is a $d \times \infty$ factor loading matrix, $\boldsymbol{\eta}_\ell \sim MVN_\infty(\mathbf{0}, \mathbf{I}_\infty)$, and $\boldsymbol{\xi}_\ell \sim MVN_c(\mathbf{0}, \boldsymbol{\Sigma}_\xi)$ with $\boldsymbol{\Sigma}_\xi = \text{diag}\{\sigma_1^2, \ldots, \sigma_c^2\}$. While it is typical to set the dimension of the latent factor $\boldsymbol{\eta}_\ell$ to be much smaller than $\boldsymbol{\epsilon}_\ell$, the approach followed in Zhu et al. (2014) is to choose a multiplicative gamma process prior for $\boldsymbol{\Lambda}$ that shrinks the elements to zero as the column index increases, thereby avoiding the issue of choosing the number of factors (see also Bhattacharya & Dunson, 2011). The overall model for the IP for a given subject can be written as

$$\mathbf{y}_\ell = \sum_{j=1}^r X_\ell^T \delta_j \mathbf{u}_j \mathbf{v}_j^T + \boldsymbol{\Lambda}\boldsymbol{\eta}_\ell + \boldsymbol{\xi}_\ell,$$

and Gaussian shrinkage priors are adopted for δ_j, $\mathbf{u}_j$, and $\mathbf{v}_j$, $j = 1, \ldots, r$. Zhu et al. (2014) present a Gibbs sampler that can be used to sample the posterior distribution and investigate a number of model selection criteria for choosing r. Their simulation studies indicate that the BIC (Bayesian Information Criterion) outperforms several other model selection criteria in determining the true rank of $\mathbf{C}$.

Overall, the use of multivariate methods over the mass univariate and voxel-wise approaches can lead to greater efficiency through the borrowing of information across related brain IPs. The approach of Wang et al. (2012a, 2012b) scales relatively well but does not provide uncertainty quantification. The Bayesian model of Greenlaw et al. (2017) addresses this issue at the expense of the greater computation required by the MCMC algorithm. As a result, the approach does

not scale as well as that of Wang et al. (2012a, 2012b) and it requires parallel computation for the selection of tuning parameters. The SRRR approach proposed by Vounou et al. (2010) allows for potentially higher dimensional datasets with an appropriate choice of the rank of the regression coefficient matrix, while the Bayesian reduced rank approach of Zhu et al. (2014) offers several advantages including uncertainty quantification and a sparse latent factor model for the covariance matrix of the response. A disadvantage of the multivariate approaches, regardless of which is chosen, is that the imaging data must be substantially reduced to a summary measure over a reasonable number of ROIs (in the hundreds at most) while the mass univariate and voxel-wise approaches can be applied to tens of thousands of voxels.

4. METHODS FOR LONGITUDINAL IMAGING GENETICS STUDIES

Longitudinal imaging genetics studies such as the ADNI study can provide insight into different rates of brain deterioration and how change in the structure of the brain over time is related to genetics. Szefer et al. (2017) have recently considered a longitudinal analysis of the ADNI database examining 75,845 SNPs in the Alzgene linkage regions and investigated associations with estimated rates of change in structural MRI measurements for 56 brain regions. Szefer et al. (2017) consider three phases of the ADNI study in their analysis: ADNI-1, ADNIGO and ADNI-2. More information on the ADNI study including information on data access is available online at http://adni.loni.usc.edu/about/. The regions considered in Szefer et al. (2017) are the same as those considered in Greenlaw et al. (2017), and are also described in Table 1 which we consider in our example analysis of the next session.

A primary innovation in the analysis of Szefer et al. (2017) is the construction from longitudinal MRI data and linear mixed models a set of subject and region specific rates of change over time. These estimated rates of change are then related to genetic markers using SCCA. Szefer et al. (2017) also use inverse probability weighting to account for the biased sampling design of the ADNI study, an aspect that has not been considered in many previous imaging genetics studies.

Let $\mathbf{y}_\ell(t) = (y_{\ell 1}(t), \ldots, y_{\ell c}(t))^T$ denote the IP summarizing the structure of the brain over c ROIs for subject ℓ, $\ell = 1, \ldots, n$, at time t, where, for the ADNI study considered by Szefer et al. (2017) $t \in \{0, 6, 12, 18, 24\}$ months following entry into the study. For the jth ROI, Szefer et al. (2017) fit the following standard linear mixed model with random intercept and slope for time

$$\begin{aligned} y_{\ell j}(t) = {} & \beta_{0j} + \beta_{1j}\mathrm{MCI} + \beta_{2j}\mathrm{AD} + \beta_{3j}t + \beta_{4j}\mathrm{MCI} \times t \\ & + \beta_{5j}\mathrm{AD} \times t + \gamma_{1\ell j} + \gamma_{2\ell j}t + \epsilon_{\ell j}(t), \end{aligned} \tag{5}$$

where AD is an indicator for AD, MCI is an indicator for mild cognitive impairment, the β terms denote fixed effects and the γ terms denote random effects. The estimated rate of change extracted from the fitted linear mixed model is $\hat{\beta}_{3j} + \hat{\beta}_{4j}\mathrm{MCI} + \hat{\beta}_{5j}\mathrm{AD} + \hat{\gamma}_{2\ell j}$, and these estimates, which are region specific, are used as the IP in the second stage of their analysis after adjusting for population stratification using multidimensional scaling. The genetic markers are also adjusted for population stratification using the principal coordinates obtained from multidimensional scaling.

A sparse linear combination of the SNP genotypes that is most associated with a linear combination of the IPs (the estimated rates of change) is obtained using SCCA. SCCA is a multivariate method for estimating maximally correlated sparse linear combinations of the columns of two multivariate data sets. The degree of sparsity in the coefficients of the genotypes is controlled by a regularization parameter, and Szefer et al. (2017) choose this parameter so that approximately 10% of the SNPs have non-zero coefficients. A bootstrap procedure is then used to estimate the relative importance of each SNP. In particular, sampling with replacement

within each disease category (MCI, AD and cognitively normal), if $\boldsymbol{\beta}_b = (\beta_{1b}, \ldots, \beta_{db})^T$ denotes the coefficient vector of the sparse linear combination of SNPs estimated from the bth bootstrap sample, Szefer et al. (2017) define the complement of the importance probability for the kth SNP as

$$\text{VIP}_k = \frac{1}{B}\sum_{b=1}^{B} I\{\beta_{kb} = 0\},$$

where B is the total number of bootstrap samples. These probabilities are then used to select important subsets of SNPs. An interesting aspect of the analysis performed by Szefer et al. (2017) is that their procedure is applied to an ADNI-1 training sample to obtain subsets of important SNPs, and the authors are then able to validate many of these priority SNPs using a validation set from ADNIGO/2.

An alternative model for longitudinal imaging genetics data has been proposed recently by Lu et al. (2017). The proposed model extends the Bayesian low rank model of Zhu et al. (2014) to the longitudinal setting. Unlike the two-stage longitudinal analysis of Szefer et al. (2017), the model of Lu et al. (2017) links the time-varying neuroimaging data directly to the genetic markers in a single model that includes the data from all ROIs. Moreover, the proposed model allows for gene−age interactions so that the genetic effects on ROI volumes can vary across time.

Letting $y_{\ell j}(t)$ denote the longitudinal imaging measure obtained from subject ℓ at ROI j and time t, the model takes the form

$$y_{\ell j}(t) = \mathbf{X}_\ell^T \boldsymbol{\beta}_j + \mu_j(t) + \mathbf{w}_\ell(t)^T \boldsymbol{\gamma}_k + \mathbf{z}_\ell(t)^T \boldsymbol{b}_{\ell j} + \epsilon_{\ell j}(t), \;\; \ell = 1, \ldots, n; j = 1, \ldots c,$$

where $\mathbf{X}_\ell$ contains the genetic markers; $\mathbf{w}_\ell(t)$ is a vector of time-varying covariates that may include interactions between genetic markers and time; $\mu_j(t)$ is an overall temporal trend for the jth ROI; and $\boldsymbol{b}_{\ell j}$ is a vector of subject specific Gaussian random effects for ROI j corresponding to covariates $\mathbf{z}_\ell(t)$. Lu et al. (2017) represent the functions $\mu_j(t)$ using penalized-splines, and as in Zhu et al. (2014) a low rank approximation is used to approximate the regression coefficient matrix. The errors $\epsilon_{\ell j}(t)$ are represented through a sparse factor model

$$\epsilon_{\ell j}(t) = \boldsymbol{\Lambda}\boldsymbol{\eta}_\ell(t) + \boldsymbol{\xi}_\ell(t)$$

with priors similar to those adopted in Zhu et al. (2014), including a multiplicative gamma process prior for $\boldsymbol{\Lambda}$. A Gibbs sampling algorithm is used to implement Bayesian inference.

Overall, methods for longitudinal imaging genetics studies are just in their infancy, with very few published articles developing statistical methods to date. We believe there is significant scope for new work in this sub-area. Regarding the methods discussed here, a primary difference in the work of Lu et al. (2017) and that proposed by Szefer et al. (2017) is that the regression model on genetic markers in the latter case is built on estimated rates of change of the ROI volumes; whereas, Lu et al. (2017) link the genetic data directly to the mean of the ROI volumes. Both methods provide useful and complementary techniques for analyzing longitudinal imaging genetics data.

5. EXAMPLE APPLICATION

We provide an example application examining an imaging genetics dataset obtained from the ADNI-1 database. The analysis presented here is considered in greater detail in Greenlaw et al. (2017); however, our objective in this case is simply to provide the reader with a simple example illustrating the use of some of the methods discussed in our review.

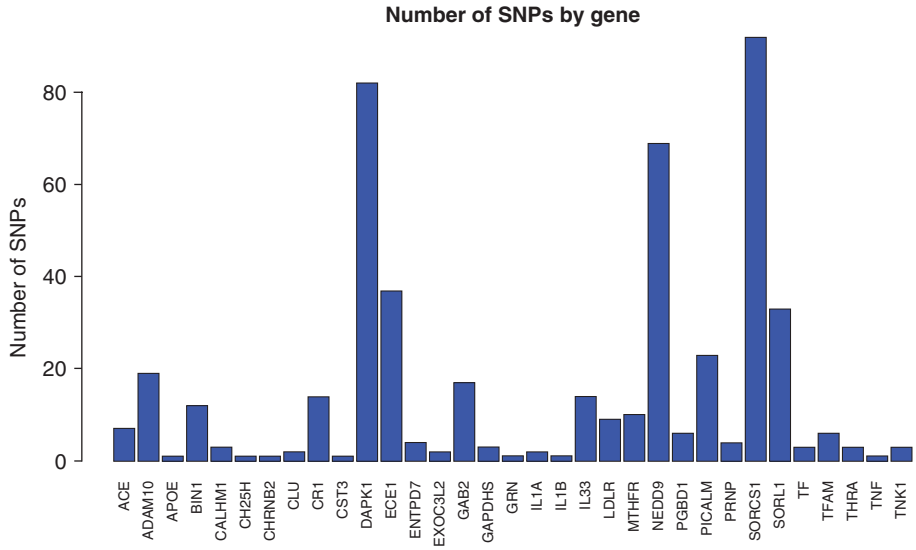

FIGURE 2: Each of the 33 genes partitioning the 486 SNPs included in the example data analysis of Section 5.

The dataset includes both genetic and structural MRI data, the latter leading to the 56 IPs presented in Table 1. The data are available for $n = 632$ subjects (179 cognitively normal, 144 Alzheimer's and 309 mild cognitive impairment), and among all possible SNPs the analysis includes only those SNPs belonging to the top 40 AD candidate genes listed on the AlzGene database as of June 10, 2010. The data presented here are queried from the most recent genome build as of December 2014, from the ADNI-1 database.

After the quality control and imputation steps have been carried out, the genetic data used for this analysis include 486 SNPs arising from 33 genes. These genes along with the number of SNPs included in our analysis from each of these genes is depicted in Figure 2. The freely available software package PLINK (Purcell et al., 2007) is used for genomic quality control. Thresholds used for SNP and subject exclusion are the same as in Wang et al. (2012a, 2012b), with the following exceptions. For SNPs, we require a more conservative genotyping call rate of at least 95% (Ge et al., 2012). For subjects, we require at least one baseline and one follow-up MRI scan and exclude multivariate outliers. Sporadically missing genotypes at SNPs in the HapMap3 reference panel (Gibbs et al., 2003) are imputed into the data using IMPUTE2 (Howie, Donnelly, & Marchini, 2009). Further details of the quality control and imputation procedure can be found in Szefer et al. (2017).

The MRI data from the ADNI-1 database are preprocessed using the FreeSurfer V4 software which conducts automated parcellation to define volumetric and cortical thickness values from the $c = 56$ brain regions of interest that are detailed in Table 1. Each of the response variables is adjusted for age, gender, education, handedness and baseline total ICV based on regression weights from healthy controls and are then scaled and centred to have zero-sample-mean and unit-sample-variance.

We fit the Bayesian model of Greenlaw et al. (2017) and also compute the group sparse multi-task regression and feature selection estimator of Wang et al. (2012a, 2012b), both of which contain $56 \times 486 = 27,216$ regression parameters. For the former approach, we select potentially important SNPs by evaluating the 95% equal-tail credible interval for each regression coefficient and select those SNPs where at least one of the associated credible intervals excludes 0. In total there are 45 SNPs and 152 regression coefficients for which this occurs. The 45

TABLE 2: The 45 SNPs selected from the Bayesian model along with corresponding phenotypes where (L), (R), (L,R) denote that the phenotypes are on the left, right, and both hemispheres respectively. SNPs also ranked among the top 45 using the Wang et al. (2012a, 2012b) estimate are listed in bold.

SNP	Gene	Phenotype ID (Hemisphere)
rs4305	ACE	LatVent (R)
rs4311	ACE	InfParietal (L,R)
		MeanPar (L,R), Precuneus (L,R)
		SupParietal (L), SupTemporal (L)
		CerebCtx (R), MeanFront (R)
		MeanSensMotor (R), MeanTemp (R)
		Postcentral (R), PostCing (R)
		Precentral (R), SupFrontal (R)
		SupParietal (R)
rs405509	APOE	AmygVol (L), CerebWM (L), Fusiform (L)
		HippVol (L), InfParietal (L,R),SupFrontal (L,R), Supramarg (L,R)
		InfTemporal (L), MeanFront (L,R), MeanLatTemp (L,R)
		MeanMedTemp (L,R), MeanPar (L,R),
		MeanSensMotor (L,R), MeanTemp (L,R)
		MidTemporal (L,R), Postcentral (L,R), Precuneus (L,R)
		SupTemporal (L,R), Precentral (R), SupParietal (R)
rs11191692	CALHM1	EntCtx (L)
rs3811450	CHRNB2	Precuneus (R)
rs9314349	CLU	Parahipp (L)
rs2025935	CR1	CerebWM (R), Fusiform (R), InfLatVent (R)
rs11141918	DAPK1	CerebCtx (R)
rs1473180	DAPK1	CerebCtx (L,R) ,EntCtx (L), Fusiform (L)
		MeanMedTemp (L), MeanTemp (L), PostCing (L)
rs17399090	DAPK1	MeanCing (R), PostCing (R)
rs3095747	DAPK1	InfLatVent (R)
rs3118846	DAPK1	InfParietal (R)
rs3124237	DAPK1	PostCing (R), Precuneus (R), SupFrontal (R)
rs4878117	DAPK1	MeanSensMotor (R), Postcentral (R)
rs212539	ECE1	PostCing (R)
rs6584307	ENTPD7	Parahipp (L)
rs11601726	GAB2	CerebWM (L), LatVent (L)
rs16924159	IL33	MeanCing (L), PostCing (L), CerebWM (R)
rs928413	IL33	InfLatVent (R)
rs1433099	LDLR	CerebCtx.adj (L), Precuneus (L,R)

TABLE 2: Continued

SNP	Gene	Phenotype ID (Hemisphere)
rs2569537	LDLR	CerebWM (L,R)
rs12209631	NEDD9	CerebCtx (L), HippVol (L,R)
rs1475345	NEDD9	Parahipp (L)
rs17496723	NEDD9	Supramarg (L)
rs2327389	NEDD9	AmygVol (L)
rs744970	NEDD9	MeanFront (L), SupFrontal (L)
rs7938033	PICALM	EntCtx (R), HippVol (R)
rs2756271	PRNP	EntCtx (L), HippVol (L,R), InfTemporal (L), Parahipp (L)
rs6107516	PRNP	MidTemporal (L,R)
rs1023024	SORCS1	MeanSensMotor (L), Precentral (L)
rs10787010	SORCS1	AmygVol (L), EntCtx (L,R)
		MeanFront (L), Fusiform (L)
		HippVol (L,R), InfLatVent (L), InfTemporal (L)
		MeanMedTemp (L,R), MeanTemp (L)
		Precentral (L), TemporalPole (R)
rs10787011	SORCS1	EntCtx (L,R), HippVol(R)
rs12248379	SORCS1	PostCing (R)
rs1269918	SORCS1	CerebCtx (L), CerebWM (L), InfLatVent (L)
rs1556758	SORCS1	SupParietal (L)
rs2149196	SORCS1	MeanSensMotor (L), Postcentral (L,R)
rs2418811	SORCS1	CerebWM (L,R), InfLatVent.adj (L)
rs10502262	SORL1	MeanCing (L), InfTemporal (R), Supramarg (R)
rs1699102	SORL1	MeanMedTemp (R), MeanTemp (R)
rs1699105	SORL1	MeanCing (L), Precuneus (L)
rs4935774	SORL1	CerebWM (L,R)
rs666004	SORL1	InfTemporal (L)
rs1568400	THRA	Precentral (L), TemporalPole (R)
rs3744805	THRA	MeanSensMotor (R), Postcentral (R), Precentral (R)
rs7219773	TNK1	MeanSensMotor (L), Precentral (L), Postcentral (R)

selected SNPs and the corresponding brain regions at which we see a potential association based on the 95% credible intervals are listed in Table 2.

Three SNPs, rs4311 from the ACE gene, rs405509 from the APOE gene and rs10787010 from the SORCS1 gene stand out as being potentially associated with the largest number of ROIs. The 95% credible intervals for the coefficients relating rs4311 to each of the $c = 56$ imaging measures are depicted in Figure 3. Both the Bayesian posterior mean and the estimator of Wang et al. (2012a, 2012b) are indicated in the figure.

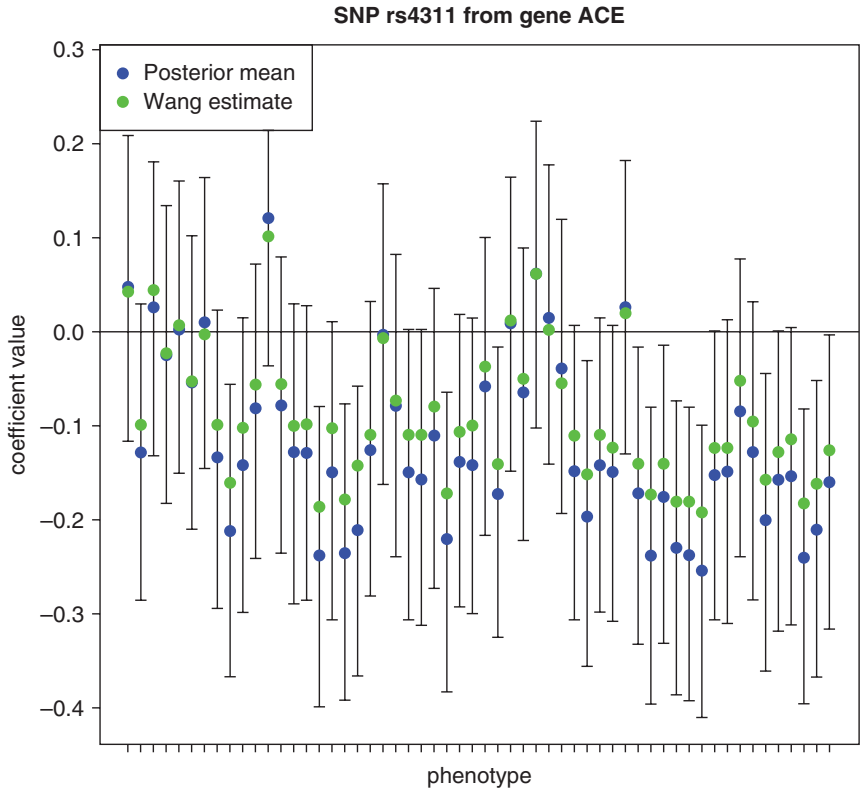

FIGURE 3: The 95% equal-tail credible intervals relating the SNP rs4311 from ACE to each of the $c = 56$ imaging phenotypes. Each imaging phenotype is represented on the x-axis with a tick mark and these are ordered in the same order as the phenotypes are listed in the rows of Table 1, first for the left hemisphere and then followed by the same phenotypes for the right hemisphere.

In the original methodology of Wang et al. (2012a, 2012b) the authors suggest ranking and selecting SNPs by constructing a SNP weight based on the point estimate $\hat{\boldsymbol{W}}$ and a sum of the absolute values of the estimated coefficients of each single SNP over all of the tasks. Doing so, the top 45 highest ranked SNPs contain 21 of the SNPs chosen using the Bayesian approach of Greenlaw et al. (2017) and these 21 SNPs are highlighted in Table 2 with bold font. The number 1 ranked (highest priority) SNP using this approach is SNP rs3026841 from gene ECE1. In Figure 4 we display the corresponding point estimates for this SNP along with the 95% credible intervals obtained from the Greenlaw et al. (2017) Bayesian approach, where again the credible intervals and point estimates are relating this SNP to each of the $c = 56$ imaging measures. *Importantly, we note that all 56 of the corresponding 95% credible intervals include the value 0.*

This result demonstrates the importance of accounting for posterior uncertainty beyond a sparse point estimate and illustrates the potential problems that may arise when estimation uncertainty is ignored, as in the approach of Wang et al. (2012a, 2012b). The methodology of Greenlaw et al. (2017) complements the estimator of Wang et al. (2012a, 2012b) by providing uncertainty quantification, and both approaches may be applied together for such analyses.

While we have focussed on uncertainty quantification using credible intervals, the posterior distribution can be summarized through posterior probabilities of the form $Pr(|W_{ij}| > \delta|\text{Data})$ for known critical value $\delta > 0$, or through kernel density estimation of the posterior density

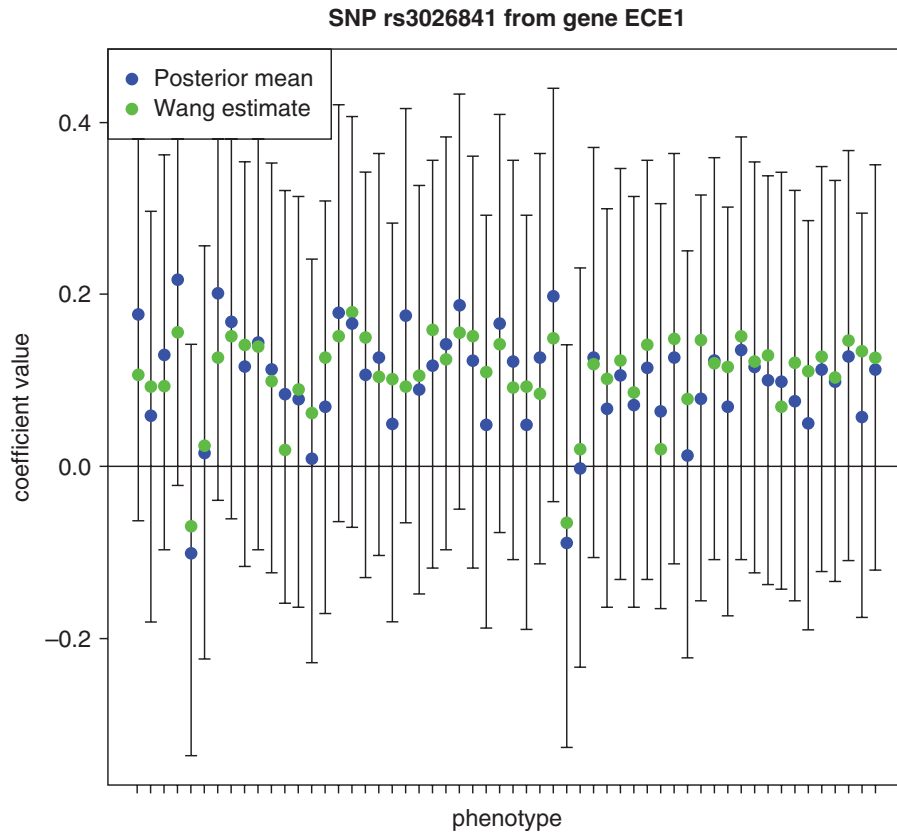

FIGURE 4: The 95% equal-tail credible intervals relating the SNP rs3026841 from ECE1 to each of the $c = 56$ imaging phenotypes. Each imaging phenotype is represented on the x-axis with a tick mark and these are ordered in the same order as the phenotypes are listed in the rows of Table 1, first for the left hemisphere and then followed by the same phenotypes for the right hemisphere.

for certain regression coefficients. In the former case, adjustments for multiplicity can be made using Bayesian FDR procedures (Morris et al., 2001).

6. DISCUSSION

Imaging genetics is an emerging discipline that is based on combining two of the most important scientific fields where statistical research has made a major impact, genetics and neuroimaging. The resulting studies provide a number of big data challenges for statistical analysis. We have reviewed a variety of approaches for the analysis of such data focussing on mass univariate and voxel-wise approaches, multivariate approaches, and methods for longitudinal studies. Figure 5 summarizes these three approaches from a graphical perspective.

One class of methods that we have not discussed in our review is the class of predictive methods for imaging genetics. In this setting the dependent variable is a condition or health outcome (such as presence of AD), and the imaging and genetic data are used for disease classification. These methods typically aim at detecting prognostic markers. Typically, regularization methods, boosting algorithms and deep learning methods are applied to such problems

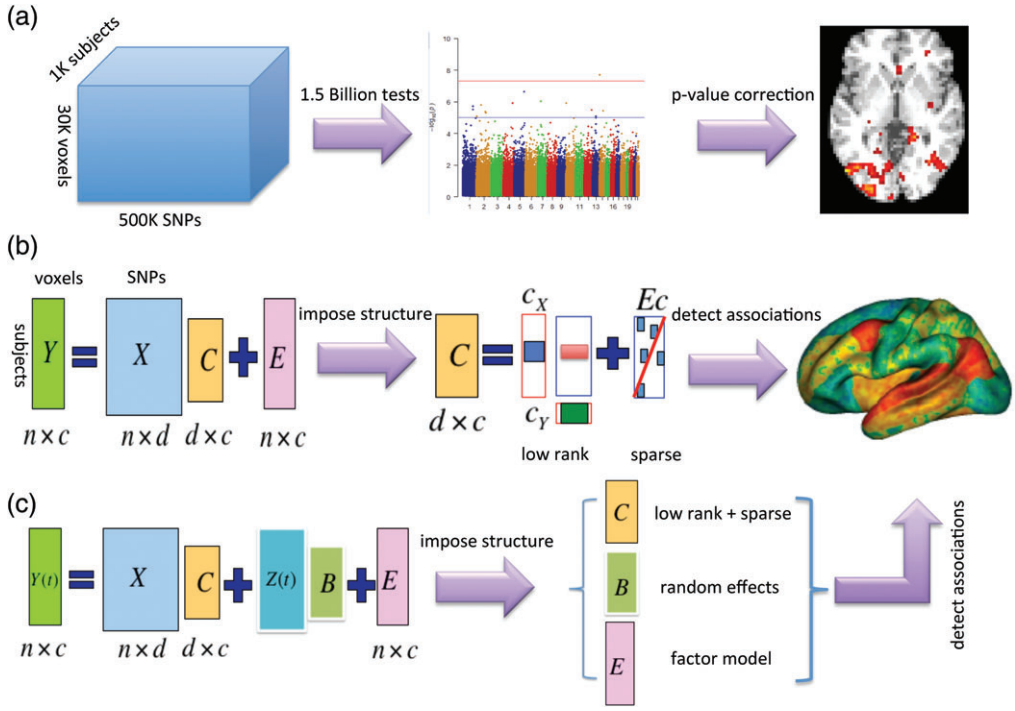

FIGURE 5: The three approaches discussed in the article summarized from a graphical perspective: (a) mass univariate and voxel-wise approaches; (b) multivariate approaches; and (c) methods for longitudinal imaging genetics studies.

(see e.g., Wang et al., 2012a, 2012b; Zhang et al., 2014). Within a Bayesian setting, a predictive model for imaging genetics with application to schizophrenia has been developed by Chekouo et al. (2016).

While our review is not an exhaustive review of existing methods for imaging genetics, our aim was to provide the reader with a sample of the existing work and a flavour of the challenges for data analysis in this area. Indeed, this is a relatively new area in statistics and there is much scope for improving the existing methods.

For example, one current avenue of interest is the extension of the methodology developed by Greenlaw et al. (2017) to accommodate a more realistic covariance structure for the IPs. One approach for doing this is through a sparse latent factor model as considered in Zhu et al. (2014) and Lu et al. (2017). An alternative approach that we are currently investigating is the use of spatial models based on Markov random fields for the regression errors. More specifically, for the data considered in Greenlaw et al. (2017), our example in Section 5, and described in Table 1, the MRI-based phenotypes will exhibit two forms of correlation: (i) spatial correlation between neighbouring ROIs on the same brain hemisphere; and (ii) correlation between corresponding measures on opposite brain hemispheres (e.g., the volume of the left hippocampus will be highly correlated with the volume of the right hippocampus).

Considering the model formulation of Greenlaw et al. (2017), we begin by rearranging the IPs so that they occur in left-right pairs in the vector $\mathbf{y}_\ell \in \mathbb{R}^c$. Let $\mathbf{y}_{\ell,i} = (y^{(L)}_{\ell,i}, y^{(R)}_{\ell,i})^T$ be the brain summary measures obtained at the ith ROI for both the left and right hemispheres. Then $\mathbf{y}_\ell = (\mathbf{y}^T_{\ell,1}, \ldots, \mathbf{y}^T_{\ell,\frac{c}{2}})^T$ is the IP for subject ℓ with the elements rearranged so that left-right pairs

are adjacent. The regression model is specified as $\mathbf{y}_\ell = \mathbf{W}^T\mathbf{x}_\ell + \boldsymbol{e}_\ell$ where a spatial model for $\boldsymbol{e}_\ell$ is based on a proper bivariate conditional autoregressive model (Gelfand & Vounatsou, 2003).

We assume $\mathbf{A}$ is an adjacency matrix $A_{ij} \in \{0, 1\}$ representing the spatial neighbourhood structure of ROIs on each hemisphere, with $\mathbf{D}_\mathbf{A} = \text{diag}\{A_{i\cdot}, i = 1, \ldots, c/2\}$. The conditional specification for the regression errors is given by

$$\boldsymbol{e}_{\ell,i}|\{\boldsymbol{e}_{\ell,j}, j \neq i\}, \rho, \Sigma \sim \text{BVN}\left(\frac{\rho}{A_{i\cdot}}\sum_{j=1}^{c/2} A_{ij}\boldsymbol{e}_{\ell,j}, \frac{1}{A_{i\cdot}}\Sigma\right), \quad i = 1, \ldots, c/2,$$

where $\rho \in [0, 1)$ characterizes spatial dependence and $\Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}} \in [-1, 1]$ characterizes the dependence in the phenotypes across opposite hemispheres of the brain.

The first level of the model can then be expressed as

$$\mathbf{y}_\ell|\mathbf{W}, \rho, \Sigma \overset{ind}{\sim} MVN_c(\mathbf{W}^T\mathbf{x}_\ell\ ,\ (\mathbf{D}_\mathbf{A} - \rho\mathbf{A})^{-1} \otimes \Sigma), \quad \ell = 1, \ldots, n$$

with higher levels of the model including the shrinkage prior for $\mathbf{W}$ specified as in Greenlaw et al. (2017) with some minor modifications. To clarify, the neighbourhood structure and resulting adjacency matrix is used in the model to represent spatial dependence between phenotypes on the *same* hemisphere of the brain, while Σ, a 2-by-2 matrix, is used to represent the correlation between the same phenotype on *opposite* hemispheres of the brain. For specification of such a model it is convenient to arrange the phenotypes into left-right pairs which is why the rearrangement is needed.

With regard to computation for this model, Greenlaw et al. (2017) and their corresponding implementation in the R package ''bgsmtr'' make use of sparse numerical linear algebra routines as the full conditional distributions for $\mathbf{W}$ have sparse precision matrices under that model. This is essential in order for our Gibbs sampling algorithm to be scalable to imaging genetics data of even moderately large size. In the proposed spatial model, so long as the adjacency matrix $\mathbf{A}$ is sparse the model structure still results in sparse precision matrices where sparse methods are required for faster computation. This is an advantage of using the Markov random field model over some other possible spatial models. The additional parameters $\rho \sim \text{Unif}(0, 1)$ and $\Sigma \sim \text{inv-Wishart}(\mathbf{S}, \nu)$ are easily added to the existing the Gibbs sampling algorithm. In addition to the use of Gibbs sampling, we are also developing a mean-field variational Bayes algorithm (see e.g., Nathoo et al., 2014) for the same model which should allow for greater scalability. We hope to report on results from this new model including a comparison of the different algorithms for Bayesian computation in a follow-up article. The MCMC and variational Bayes algorithms for fitting the new spatial model are available in the current release of the ''bgsmtr'' R package.

ACKNOWLEDGEMENTS

Linglong Kong and Farouk Nathoo are supported by funding from the Natural Sciences and Engineering Research Council of Canada and a CANSSI collaborative research team on neuroimaging data analysis. Farouk Nathoo holds a Tier II Canada Research Chair in Biostatistics for Spatial and High-Dimensional Data. Dr. Zhu's work was partially supported by NIH grants MH086633 and MH092335, NSF grants SES-1357666 and DMS-1407655, a grant from the Cancer Prevention Research Institute of Texas, and the endowed Bao-Shan Jing Professorship in Diagnostic Imaging. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012).

REFERENCES

Ashburner, J., Good, C., & Friston, K. J. (2000). Tensor based morphometry. *NeuroImage*, 11, S465.

Argyriou, A., Evgeniou, T., & Pontil, M. (2007). Multi-task feature learning. In *Advances in Neural Information Processing Systems* Cambridge: MIT Press; 41–48.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

Bhattacharya, A. & Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika*, 98, 291–306.

Cannon, T. D. & Keller, M. C. (2006). Endophenotypes in the genetic analyses of mental disorders. *Annual Review of Clinical Psychology*, 40, 267–290.

Chekouo, T., Stingo, F. C., Guindani, M., & Do, K. -A. (2016). A Bayesian predictive model for imaging genetics with application to schizophrenia. *The Annals of Applied Statistics*, 10, 1547–1571.

Friston, K. (2009). Modalities, modes, and models in functional neuroimaging. *Science*, 326, 399–403.

Ge, T., Feng, J., Hibar, D. P., Thompson, P. M., & Nichols, T. E. (2012). Increasing power for voxel-wise genome-wide association studies: The random field theory, least square kernel machines and fast permutation procedures. *NeuroImage*, 63, 858–873.

Ge, T., Nichols, T. E., Ghosh, D., Mormino, E. C., Smoller, J. W., Sabuncu, M. R., & Alzheimer's Disease Neuroimaging Initiative. (2015). A kernel machine method for detecting effects of interaction between multidimensional variable sets: An imaging genetics application. *NeuroImage*, 109, 505–514.

Gelfand, A. E. & Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4, 11–15.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24, 997–1016.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F. L., Yang, H. M., Ch'ang, L. -Y., & the International HapMap Consortium. (2003). The international HapMap project. *Nature*, 426, 789–796.

Greenlaw, K., Szefer, E., Graham, J., Lesperance, M., Nathoo, F. S., & Alzheimer's Disease Neuroimaging Initiative. (2017). A Bayesian group sparse multi-task regression model for imaging genetics. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btx215.

Hibar, D. P., Stein, J. L., Renteria, M. E., Arias-Vasquez, A., Desrivières, S., Jahanshad, N., Toro, R., Wittfeld, K., Abramovic, L., Andersson, M., & Aribisala, B. S. (2015). Common genetic variants influence human subcortical brain structures. *Nature*, 520, 224.

Hibar, D. P., Stein, J. L., Kohannim, O., Jahanshad, N., Saykin, A. J., Shen, L., Kim, S., Pankratz, N., Foroud, T., Huentelman, M. J., Potkin, S. G., Jack, C. R. Jr., Weiner, M. W., Toga, A. W., Thompson, P. M., & Alzheimer's Disease Neuroimaging Initiative. (2011). Voxel-wise gene-wide association study (vGeneWAS): Multivariate gene-based association testing in 731 elderly subjects. *NeuroImage*, 56, 1875–1891.

Hibar, P., Stein, J., Renteria, M., Arias-Vasquez, A., Desrivieres, S., Jahanshad, A., & et al. (2015). Common genetic variants influence human subcortical brain structure. *Nature*, 520, 224–229.

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5, e1000529.

Huang, M., Nichols, T., Huang, C., Yu, Y., Lu, Z., Knickmeyer, R. C., Feng, Q., Zhu, H., & Alzheimer's Disease Neuroimaging Initiative. (2015). FVGWAS: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. *NeuroImage*, 118, 613–627.

Huang, C., Thompso, P., Wang, Y., Yu, Y., Zhang, J., Kong, D., Cole, R., Knickmeyer, R. C., Zhu, H., & Alzheimer's Disease Neuroimaging Initiative. (2017). FGWAS: Functional genome wide association analysis. *NeuroImage*, 159, 107–121.

Kotz, S., Kozubowski, T., & Podgorski, K. (2012). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Springer Science & Business Media.

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5, 369–411.

Liu, D., Lin, X., & Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63, 1079–1088.

Lu, Z. H., Khondker, Z., Ibrahim, J. G., Wang, Y., Zhu, H., & Alzheimer's Disease Neuroimaging Initiative. (2017). Bayesian longitudinal low-rank regression models for imaging genetic data from longitudinal studies. *NeuroImage*, 149, 305–322.

Marenco, S. & Radulescu, E. (2010). Imaging genetics of structural brain connectivity and neural integrity markers. *NeuroImage*, 53, 848–856.

Miller, M. I. & Younes, L. (2001). Group actions, homeomorphisms, and matching: A general framework. *International Journal of Computer Vision*, 41, 61–84.

Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A., & Coombes, K. R. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics*, 64, 479–489.

Munafo, M. R. & Flint, J. (2011). Dissecting the genetic architecture of human personality. *Trends in Cognitive Sciences*, 15, 395–400.

Nathoo, F. S., Greenlaw, K., & Lesperance, M. (2016). Regularization parameter selection for a Bayesian group sparse multi-task regression model with application to imaging genomics. In *International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, IEEE, New York, p. 1–4.

Nathoo, F. S., Babul, A., Moiseev, A., Virji-Babul, N., & Beg, M. F. (2014). A variational Bayes spatiotemporal model for electromagnetic brain mapping. *Biometrics*, 70, 132–143.

Park, T. & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103, 681–686.

Peper, J., Brouwer, R., Boomsma, D., Kahn, R., & Pol, H. (2007). Genetic influences on human brain structure: A review of brain imaging studies in twins. *Human Brain Mapping*, 28, 464–473.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81, 559–575.

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., & Alzheimer's Disease Neuroimaging Initiative. (2010). Voxel-wise genome-wide association study (vGWAS). *NeuroImage*, 53, 1160–1174.

Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M. E., Hopson, R., Jackson, C., Keefe, J., Riley, M., Mentch, F. D., Sleiman, P., Verma, R., Davatzikos, C., Hakonarson, H., Gur, R. C., & Gur, R. E. (2014). Neuroimaging of the Philadelphia neurodevelopmental cohort. *NeuroImage*, 86, 544–553.

Stingo, F. C., Guindani, M., Vannucci, M., & Calhoun, V. D. (2013). An integrative Bayesian modeling approach to imaging genetics. *Journal of the American Statistical Association*, 108, 876–891.

Szefer, E., Lu, D., Nathoo, F., Beg, M. F., Graham, J., & Alzheimer's Disease Neuroimaging Initiative. (2017). Multivariate association between single-nucleotide polymorphisms in Alzgene linkage regions and structural changes in the brain: Discovery, refinement and validation. *Statistical Applications in Genetics and Molecular Biology*, 16, 349–365.

Thompson, P. M., Ge, T., Glahn, D. C., Jahanshad, N., & Nichols, T. E. (2013). Genetics of the connectome. *NeuroImage*, 80, 475–488.

Vounou, M., Nichols, T. E., Montana, G., & Alzheimer's Disease Neuroimaging Initiative. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *NeuroImage*, 53, 1147–1159.

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., Saykin, A. J., Shen, L., & Alzheimer's Disease Neuroimaging Initiative. (2012a). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the ADNI cohort. *Bioinformatics*, 28, 229–237.

Wang, H., Nie, F., Huang, H., Risacher, S. L., Saykin, A. J., Li, S., & Alzheimer's Disease Neuroimaging Initiative. (2012b). Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, 28, 127–136.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, 11, 3571–3594.

Worsley, K. (2002). Non-stationary FWHM and its effect on statistical inference of fMRI data. *NeuroImage*, 15, 779–790.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4, 58–73.

Xu, D., Mori, S., Shen, D., van Zijl, P., & Davatzikos, C. (2003). Spatial normalization of diffusion tensor fields. *Magnetic Resonance in Medicine*, 50, 175–182.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67.

Zhang, Z., Huang, H., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. (2014). Integrative analysis of multi-dimensional imaging genomics data for Alzheimer's disease prediction. *Frontiers in Aging Neuroscience*, 6, A260.

Zhu, H., Gu, M., & Peterson, B. (2007). Maximum likelihood from spatial random effects models via the stochastic approximation expectation maximization algorithm. *Statistics and Computing*, 15, 163–177.

Zhu, H., Khondker, Z., Lu, Z., & Ibrahim, J. G. (2014). Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association*, 109, 977–990.

Received 24 July 2017
Accepted 08 October 2018