NeuroImage 149 (2017) 305–322

Contents lists available at ScienceDirect

NeuroImage

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

Bayesian longitudinal low-rank regression models for imaging genetic data from longitudinal studies

Zhao-Hua Lu[a], Zakaria Khondker[b], Joseph G. Ibrahim[b], Yue Wang[b], Hongtu Zhu[c,*], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA
[b] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[c] Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, USA

ARTICLE INFO

Keywords:
Genetic variants
Longitudinal imaging phenotypes
Low-rank regression
Markov chain Monte Carlo
Spatiotemporal correlation

ABSTRACT

To perform a joint analysis of multivariate neuroimaging phenotypes and candidate genetic markers obtained from longitudinal studies, we develop a Bayesian longitudinal low-rank regression (L2R2) model. The L2R2 model integrates three key methodologies: a low-rank matrix for approximating the high-dimensional regression coefficient matrices corresponding to the genetic main effects and their interactions with time, penalized splines for characterizing the overall time effect, and a sparse factor analysis model coupled with random effects for capturing within-subject spatio-temporal correlations of longitudinal phenotypes. Posterior computation proceeds via an efficient Markov chain Monte Carlo algorithm. Simulations show that the L2R2 model outperforms several other competing methods. We apply the L2R2 model to investigate the effect of single nucleotide polymorphisms (SNPs) on the top 10 and top 40 previously reported Alzheimer disease-associated genes. We also identify associations between the interactions of these SNPs with patient age and the tissue volumes of 93 regions of interest from patients' brain images obtained from the Alzheimer's Disease Neuroimaging Initiative.

1. Introduction

Many longitudinal neuroimaging studies concomitantly collect genetic and recurrent imaging data to track individual changes in brain structure and function over time. Several neurodegenerative disorders, including Alzheimer disease (AD), are hypothesized to occur from abnormal development of the brain, which may be caused by the additive and/or interactive effects of various risk genes and environmental risk factors, each contributing small individual effects. Thus, recurrent neuroimaging measures may lead to discoveries of the genetic pathways and the causal genes associated with the specific brain changes underlying such neurodegenerative disorders (Scharinger et al., 2010; Paus, 2010; Peper et al., 2007; Chiang et al., 2011a, 2011b; Saykin et al., 2015).

A standard statistical method used in longitudinal imaging and genetics studies is the massive marginal association (MMA) framework (Li et al., 2013; Zhang et al., 2014; Guillaume et al., 2014; Hibar, 2011; Shen et al., 2010; Bernal-Rusiel et al., 2013; Zhang et al., 2014). This approach repeatedly fits a linear mixed effects model (or generalized estimating equations) for paired imaging phenotypes and genetic markers. Because the MMA framework entails numerous comparisons, it can detect only phenotype-marker pairs with extremely strong signals.

Several attempts have been made to more precisely investigate the effect of multiple genotypes on longitudinal phenotypes. Chen and Wang (2011) proposed functional mixed-effect models with penalized splines and varying coefficients, but they focused on small number of predictors and number of response variables in a low-dimensional setting. Wang et al. (2011) used a sparse multitask regression to examine the association between genetic markers and longitudinal neuroimaging phenotypes. However, their model focused on subjects with the same number of repeated measures and ignored the spatio-temporal correlations of imaging phenotypes. Therefore, the multitask regression model may lead to loss of statistical power to detect phenotype-marker pairs with moderate to weak signals. Vounou et al. (2011) and Silver et al. (2012) proposed that a sparse reduced-rank

* Corresponding author.
E-mail address: hzhu5@mdanderson.org (H. Zhu).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). 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 listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

http://dx.doi.org/10.1016/j.neuroimage.2017.01.052
Received 18 October 2016; Accepted 22 January 2017
Available online 29 January 2017
1053-8119/

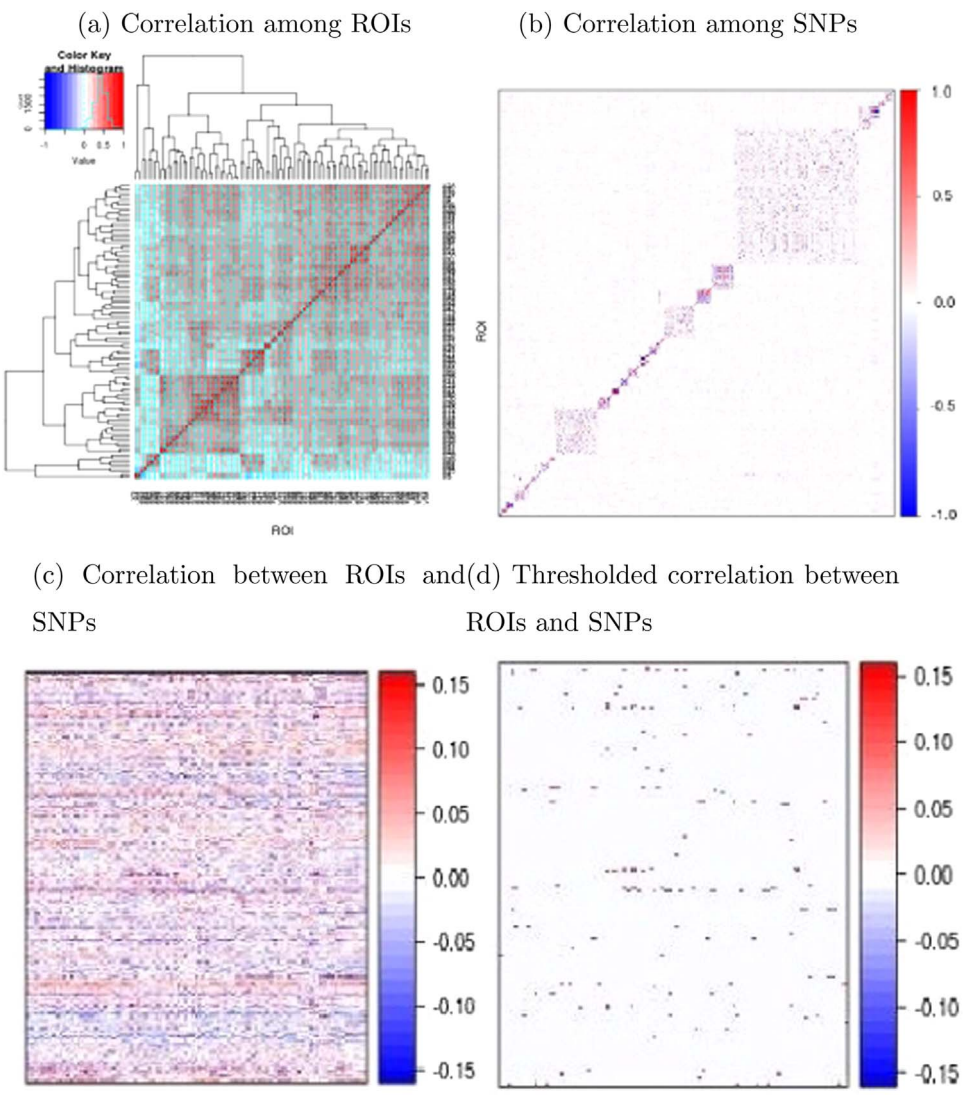

Fig. 1. The characteristics of the ROIs and SNPs from the ADNI data set and their association.

regression model using penalized regression can detect the main genetic effects on longitudinal phenotypes. They, however, did not account for the spatiotemporal association among the longitudinal phenotypes, which is important for estimation and prediction accuracy. Moreover, none of these studies explored SNP-age interactions, which can reveal dynamic genetic effects on phenotypes.

Several important statistical concerns are associated with the joint analysis of neuroimaging phenotypes and a set of candidate genotypes obtained from longitudinal imaging and genetic studies. First, the number of regression coefficients can be much larger than the sample size, denoted as N. Specifically, let d and p be the dimension of the responses and the number of covariates, respectively. Fitting a multivariate linear mixed effects model usually requires estimating a $d \times p$ matrix of regression coefficients, which can be much larger than N, even for moderately high d and p. Second, as illustrated in Fig. 1a, to improve prediction accuracy (Breiman and Friedman, 1997), it is critically important to account for unstructured, within-subject spatial correlations among multivariate neuroimaging phenotypes. Third, as illustrated in Fig. 2a, to improve both estimation and prediction accuracy, it is also important to account for within-subject temporal correlation. Fourth, as shown in Fig. 2b,c, the temporal growth pattern varies across regions of interest (ROIs) in the brain. Accounting for the overall longitudinal change of ROIs is required to increase the detection power of the genetic effects. Fifth, as shown in Fig. 2b, the genetic effects on ROI volumes can vary across time.

Here, we have developed a Bayesian longitudinal low-rank regression (L2R2) model for the joint analysis of high-dimensional longitudinal responses and covariates. We integrated multiple robust methods to explicitly address the new challenges described previously. Our study has four major methodological contributions that were previously undescribed:

1. To the best of our knowledge, L2R2 is the first model of its kind for jointly analyzing high-dimensional longitudinal responses and covariates, although several approaches have been used for high-dimensional responses and covariates in cross-sectional studies (Rothman et al., 2010; Vounou et al., 2010; Zhu et al., 2014). The L2R2 model also provides a set of standard inference tools (e.g. standard deviation) for determining various unknown parameters. Zipunnikov et al. (2014) proposed a functional principle components analysis for high dimensional ($>10,000$) longitudinal responses, where the intercepts and slopes of time for all voxels were modeled by a few basis functions. However, their methods cannot handle high-dimensional responses and covariates simultaneously because the dimension of the covariance matrix that requires time-consuming spectral decomposition equals the product of the dimensions of the responses and covariates.