NeuroImage 144 (2017) 35–57

Contents lists available at ScienceDirect

NeuroImage

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

Generalized reduced rank latent factor regression for high dimensional tensor fields, and neuroimaging-genetic applications

Chenyang Tao[a,b], Thomas E. Nichols[c], Xue Hua[d], Christopher R.K. Ching[d,e], Edmund T. Rolls[b,f], Paul M. Thompson[d,g], Jianfeng Feng[a,b,h,*], The Alzheimer's Disease Neuroimaging Initiative[1]

[a] Centre for Computational Systems Biology and School of Mathematical Sciences, Fudan University, Shanghai, PR China
[b] Department of Computer Science, Warwick University, Coventry, UK
[c] Department of Statistics, University of Warwick, Coventry, UK
[d] Imaging Genetics Center, Institute for Neuroimaging & Informatics, University of Southern California, Los Angeles, CA, USA
[e] Interdepartmental Neuroscience Graduate Program, UCLA School of Medicine, Los Angeles, CA, USA
[f] Oxford Centre for Computational Neuroscience, Oxford, UK
[g] Departments of Neurology, Psychiatry, Radiology, Engineering, Pediatrics, and Ophthalmology, USC, Los Angeles, CA, USA
[h] School of Life Science and the Collaborative Innovation Center for Brain Science, Fudan University, Shanghai 200433, PR China

ARTICLE INFO

Keywords:
Dimension reduction
Generalized linear model
High dimensional tensor field
Latent factor
Least squares kernel machines
Nuclear norm regularization
Reduced rank regression
Riemannian manifold

ABSTRACT

We propose a generalized reduced rank latent factor regression model (GRRLF) for the analysis of tensor field responses and high dimensional covariates. The model is motivated by the need from imaging-genetic studies to identify genetic variants that are associated with brain imaging phenotypes, often in the form of high dimensional tensor fields. GRRLF identifies from the structure in the data the effective dimensionality of the data, and then jointly performs dimension reduction of the covariates, dynamic identification of latent factors, and nonparametric estimation of both covariate and latent response fields. After accounting for the latent and covariate effects, GRLLF performs a nonparametric test on the remaining factor of interest. GRRLF provides a better factorization of the signals compared with common solutions, and is less susceptible to overfitting because it exploits the effective dimensionality. The generality and the flexibility of GRRLF also allow various statistical models to be handled in a unified framework and solutions can be efficiently computed. Within the field of neuroimaging, it improves the sensitivity for weak signals and is a promising alternative to existing approaches. The operation of the framework is demonstrated with both synthetic datasets and a real-world neuroimaging example in which the effects of a set of genes on the structure of the brain at the voxel level were measured, and the results compared favorably with those from existing approaches.

1. Introduction

The past decade has witnessed the dawn of the big data era. Advances in technologies in areas such as genomics and medical imaging, among others, have presented us with an unprecedentedly large volume of data characterized by high dimensionality. This not only brings opportunities but also poses new challenges to scientific research. Neuroimaging-genetics, one of the burgeoning interdisciplinary fields emerging in this new era, aims at understanding how the genetic makeup affects the structure and function of the human brain and has received increasing interest in recent years.

Starting with candidate gene and candidate phenotype studies, imaging-genetic methods have made significant progress over the years (Thompson et al., 2013; Liu and Calhoun, 2014; Poline et al., 2015). Different strategies have been implemented to combine the genetic and neuroimaging information, producing many promising results (Hibar et al., 2015; Richiardi et al., 2015; Jia et al., 2016). Using a few summary variables of the brain features is the most popular approach in the literature (Joyner et al., 2009; Potkin et al., 2009; Vounou et al., 2010); voxel-wise and genome-wide association approaches offer a more holistic perspective and are used in exploratory studies (Hibar et al., 2011; Vounou et al., 2012); multivariate analyses have also been used to capture the epistatic and pleiotropic interactions, therefore boosting the overall sensitivity (Hardoon et al., 2009; Ge at al.,

* Corresponding author at: Centre for Computational Systems Biology and School of Mathematical Sciences, Fudan University, Shanghai, PR China.
E-mail address: jianfeng64@gmail.com (J. Feng).

[1] Data used in preparation of this article were obtained from the Alzheimers 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.2016.08.027
Received 26 December 2015; Accepted 14 August 2016
Available online 22 September 2016

2015a,b). Apart from the population studies, family-based studies offer additional insights on the genetic heritability (Ganjgahi et al., 2015). Recently, a few probabilistic approaches have been proposed to jointly model the interactions between genetic factors, brain endophenotypes and behavior phenotypes (Batmanghelich et al., 2013, Stingo et al., 2013), and some Bayesian methods originally developed for eQTL studies can also be applied to imaging-genetic problems (Zhang and Liu, 2007; Jiang and Liu, 2015).

The trend in imaging-genetics is to embrace brain-wide genome-wide association studies with multivariate predictors and responses, but this is challenged by the combinatorial complexity of the problem. For example, the probabilistic formulations do not scale well with dimensionality; and standard brute force massive univariate approaches (Stein et al., 2010a; Vounou et al., 2012) treat each voxel and predictor as independent units and compute pairwise significance, and the loss of spatial information and the colossal multiple comparison corrections involved have high costs in terms of sensitivity (Hua et al., 2015). Various attempts have been made to remedy this. Some approaches involve dimension reduction techniques, which either first embed genetic factors onto some lower dimensional space using methods such as principal component analysis (PCA) before subsequent analyses (Hibar et al., 2011), or jointly project genetic factors and imaging traits by methods such as parallel independent component analysis (pICA), canonical correlation analysis (CCA) and partial least square (PLS) (Liu et al., 2009; Le Floch et al., 2012, 2013). These methods often lack model interpretability. Other popular approaches enforce penalties or constraints to regularize the solutions, for example (group) sparsity or rank constraints (Wang et al., 2012a,b; Vounou et al., 2012; Lin et al., 2015; Huang et al., 2015). But they are usually difficult to compute and the significance of the findings cannot be directly evaluated.

One path towards more efficient estimation for brain-wide association, both in the statistical and computational sense, is to exploit the inherent spatial structure from the neuroimaging data. Two prominent examples in this direction are *random field theory* based methods (Worsley et al., 1996, Penny et al., 2011; Ge et al., 2012) and *functional* based methods (Wahba, 1990; Ramsay and Silverman, 2005; Reiss and Ogden, 2010) where the smoothness of the data is considered. Random field methods are established as the core inferential tool in neuroimaging studies. These methods correct the statistical thresholds based on the smoothness estimated from the images, resulting in increased sensitivity. Functional based methods explicitly use smooth fields parametrized by smooth basis functions in the model, thereby regularizing the solution and simplifying the estimation at the same time. Related to functional methods are tensor-based methods (Zhou et al., 2013; Li, 2014) and wavelet-based methods (Van De Ville et al., 2007; Wang et al., 2014), where either low rank tensor factorization or a wavelet basis is used to approximate the spatial field of interest.

Long overlooked in neuroimaging studies, including imaging-genetics, is the influence from unobservable latent factors (Bhattacharya and Dunson et al., 2011; Montagna et al., 2012). An illustrative cartoon is presented in Fig. 1 for a typical neuroimaging-genetic case, in which the effect of interest is usually small compared with the total variance. This is known as low *signal to noise ratio* (SNR). Large-scale multi-center collaborations have become a common practice in the neuroimaging community (Jack et al., 2008; Michael et al., 2012; Van Essen et al., 2013; Thompson et al., 2014) and increasing numbers of researchers are starting to pool data from different sources. The heterogeneity of the data introduces large unexplained variance originating from population stratification or cryptic relatedness, for example genetic background, medical history, traumatic experiences and environmental impacts. Such variance aggregates the SNR issue and confuses the estimation procedures if unaccounted for. However these confounding factors are usually difficult or costly to quantify, and therefore they are hidden from the data analysis in most, if not all, studies.

To see how the latent factor-induced variance undermines the power of statistical procedures, let us take the most commonly used least squares regression as an example. Assume the model $Y = X\beta + L + E$, where Y is the response, X is the predictor of interest, β is the regression coefficient, L is the unobservable latent factor and E is the noise term. In the absence of knowledge regarding L, the alternative model $Y = X\widetilde{\beta} + \widetilde{E}$ is estimated instead, where $\widetilde{E} = L + E$. Assuming independence between X, L and E, we have $\mathrm{var}[\widetilde{E}] = \mathrm{var}[L] + \mathrm{var}[E]$, where $\mathrm{var}[\cdot]$ measures the variance. Denote $\widehat{\beta}$ the oracle estimator where the true model is fit with the knowledge of L and $\widehat{\widetilde{\beta}}$ the estimator for the alternative model, the asymptotic theory of least square estimator tells us $\widehat{\beta} \sim \mathcal{N}(\beta, \mathrm{var}[E](X'X)^{-1})$ and $\widehat{\widetilde{\beta}} \sim \mathcal{N}(\beta, \mathrm{var}[\widetilde{E}](X'X)^{-1})$ as the sample size goes to infinity, that is to say $\widehat{\widetilde{\beta}}$ is more variable than $\widehat{\beta}$ and converges slowly to the population mean. See Fig. 2 for a graphical illustration.

Solutions have been proposed to alleviate the loss of statistical efficiency caused by latent factors. In Zhu et al. (2014) the authors propose to dynamically estimate the latent factors from the observed data. However this approach is based on *Markov chain Monte-Carlo* (MCMC) sampling, and therefore the computational cost is prohibitive for high dimensional tensor field applications. In the eQTL literature, several methods that explicitly account for the hidden determinants have been developed. Following a Bayesian formulation, Stegle et al. (2010) factors out the hidden effect; Fusi et al. (2012), however, computes the ML estimate of hidden factors by marginalizing out the regression coefficients and then using the estimated hidden factors to construct certain covariance matrices for subsequent analyses. These studies are not concerned with the spatial structure and the inherent dimensionality of the model, and the results depend on the choice of parameters for the prior distributions. Additionally, these studies consider latent effect as “variance of no interest”, but as we will see in later sections, the latent structure also contains vital information and therefore should not be simply disregarded as unwanted variance.

In this article, we formulate a new generalized reduced rank latent factor regression model (GRRLF) for high dimensional tensor fields. Our method exploits the spatial structure of the neuroimaging data and the low rank structure of the regression coefficient matrix, which computes the effective covariate space, improves the generalization performance and leads to efficient estimation. The model works for general tensor field responses which include a wide range of imaging modalities, i. e. MRI, EEG, PET, etc. Although motivated by imaging-genetic applications, the proposed GRRLF is thus widely applicable to almost all types of neuroimaging studies. The estimation is carried out via minimizing a properly defined loss function, which includes *maximum likelihood estimation* (MLE) and *penalized likelihood estimation* (PLE) as special cases.

The contributions of this paper are four-fold. Firstly, we introduce field-constrained latent factor estimation for high dimensional tensor field regression analysis. It efficiently explains the covariance structure in the data caused by the hidden structures. Secondly, our model integrates dimension reduction, that not only improves the statistical efficiency but also facilitates model interpretability. Thirdly, we provide several implementations to efficiently compute the solution under constraints, including *Riemannian manifold optimization* (Absil et al., 2009) and *nuclear norm regularization* which are both based on manifold optimization. We highlight the flexibility of using manifold optimization to formulate neuroimaging problems, which can lead to further interesting applications. Lastly, we present an efficient kernel approach for brain-wide genome-wide association studies under the GRRLF framework and apply it to the ADNI dataset. Empirical results provide evidence that the kernel GRRLF approach is capable of capturing the interactions that can be missed in conventional studies.

The rest of the paper is organized as follows. In Section 2, we detail the model formulation and estimation. In Section 3, the proposed method is evaluated with both synthetic and real-world examples and