Computational Statistics and Data Analysis 105 (2017) 59–75

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Fitting large-scale structured additive regression models using Krylov subspace methods[☆,☆☆]

Paul Schmidt[a,b,c,*], Mark Mühlau[b,c], Volker Schmid[a]

[a] Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany
[b] Department of Neurology, Technische Universität München, Munich, Germany
[c] TUM-Neuroimaging Center, Technische Universität München, Munich, Germany

ARTICLE INFO

Article history:
Received 29 December 2015
Received in revised form 8 July 2016
Accepted 8 July 2016
Available online 21 July 2016

Keywords:
Markov chain Monte Carlo
Krylov subspace methods
Lanczos algorithm
Structured additive regression
Gaussian Markov random field
Image analysis

ABSTRACT

Fitting regression models can be challenging when regression coefficients are high-dimensional. Especially when large spatial or temporal effects need to be taken into account the limits of computational capacities of normal working stations are reached quickly. The analysis of images with several million pixels, where each pixel value can be seen as an observation on a new spatial location, represent such a situation. A Markov chain Monte Carlo (MCMC) framework for the applied statistician is presented that allows to fit models with millions of parameters with only low to moderate computational requirements. The method combines a modified sampling scheme with novel accomplishments in iterative methods for sparse linear systems. This way a solution is given that eliminates potential computational burdens such as calculating the log-determinant of massive precision matrices and sampling from high-dimensional Gaussian distributions. In an extensive simulation study with models of moderate size it is shown that this approach gives results that are in perfect agreement with state-of-the-art methods for fitting structured additive regression models. Furthermore, the method is applied to two real world examples from the field of medical imaging.

1. Introduction

Structured additive regression (STAR) models (Fahrmeir et al., 2004) are an important tool for the applied statistician. They allow adequate modeling of temporal, spatial and spatio-temporal effects as well as non-linear effects for continuous covariates (Lang and Brezger, 2004). As the ability to store and process large amounts of data has been increased over the past years the need to fit models with high-dimensional regression coefficients has increased as well. For example, in spatial statistics available spatial information can be included as a latent Gaussian Markov random field (GMRF, Rue and Held, 2005) using hierarchical models (Banerjee et al., 2014) which may lead to models with hundreds of thousands of

☆ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). 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.

☆☆ The Matlab code used in this paper is provided as supplementary material.

* Corresponding author at: Department of Neurology, Technische Universität München, Munich, Germany.
E-mail address: paul.schmidt.mail@gmail.com (P. Schmidt).

http://dx.doi.org/10.1016/j.csda.2016.07.006
0167-9473/

parameters (Eidsvik et al., 2012). Further examples can be found in the analysis of biological and medical images (Tabelow et al., 2011). The fixed lattice structure of these two- or three-dimensional images imposes a natural dependence structure between adjacent pixels or voxels. To account for this information, hierarchical models can be used that have regression coefficients of the size of the image dimension which can be up to several million elements (Gössl et al., 2000; Schmid et al., 2006).

Even though computational equipment constantly improves over time there will always be situations in which most ordinary working stations are on the edge of their capacities if the dimension of the problem exceeds a certain threshold. To deal with this problem one can either increase the computational capacities or modify the model that needs to be fitted until it meets the existing infrastructure. While the former solution is quite expensive and not always possible, much research has been focused on methods that deal with the latter aspect. Examples include approaches based on an approximate likelihood (Vecchia, 1988; Stein et al., 2004; Fuentes, 2007) as well as data reductions methods (Banerjee et al., 2008; Eidsvik et al., 2012). Even though these approaches sometimes provide us with the only useful solution, they still have the disadvantage that they either yield only approximate solutions or do not take into account all available information. In contrast, only few approaches exist that are able to use all information and yet give satisfactory results. The most promising approach we found was the one recently proposed by Wood et al. (2015). They presented a relatively fast method for fitting generalized additive models to large data sets by utilizing iterative updating schemes for the factorization of model matrices. However, this approach still fails with respect to computational requirements when applied to situations we are facing every day (see the applications below).

With respect to Bayesian inference on STAR models, both deterministic and MCMC based techniques have been published over the last decades. Examples for deterministic approaches are the empirical Bayes approaches in BayesX (Brezger et al., 2005; Umlauf et al., 2015) and the integrated nested Laplace approximation (INLA) approach (Rue et al., 2009). While these methods produce fast and accurate results for models up to a certain size they are currently not suited to handle models with high-dimensional parameters. This is mainly due to the fact that a joint (prior and full conditional) distribution for all regression coefficients needs to be set up in order to correctly account for dependencies between regression coefficients. In the presence of high dimensional coefficients it can be difficult to handle this joint distribution on ordinary working stations. On the other hand, MCMC based inference allows us to work on each parameter separately, thus representing an appropriate basis for large-scale problems.

Most MCMC schemes for this type of models require the calculation of the log-determinant of a precision matrix $\boldsymbol{Q}$. For moderate sized problems this is done by first factorizing $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}'$ using the Cholesky decomposition and then calculating $\log|\boldsymbol{Q}| = 2\sum_i \log L_{ii}$ (Rue, 2001). Although the lower triangular matrix $\boldsymbol{L}$ can be computed efficiently by making use of the sparsity structure of $\boldsymbol{Q}$ it still is a computational bottleneck or even impossible if $\boldsymbol{Q}$ is too large. Here we present a MCMC sampling scheme that offers the possibility to avoid the calculation of the log-determinant completely by the cost of higher acceptance rates.

The second bottleneck within most MCMC algorithms in the presence of high-dimensional regression coefficients is to sample from the proposal distribution which requires to obtain a sample $\boldsymbol{x}$ from $\mathrm{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1})$. Direct sampling usually utilizes the Cholesky decomposition and solves $\boldsymbol{L}'\boldsymbol{x} = \boldsymbol{z}$ where $\boldsymbol{z}$ is a vector of standard normal random variables. If this factorization is not possible alternative methods must be applied. Blocking (Roberts and Sahu, 1997) and sequential strategies (Winkler, 2003) as well as the conjugate priors approach by Knorr-Held (1999) represent such alternatives. While these blocking approaches are easy to implement, it turns out that the choice of blocks and the visiting order is a non-trivial subject and has a major influence on the performance of the sampler (Winkler, 2003). In our experience it cannot even be guaranteed that every choice of blocks will lead to the correct stationary distribution of the MCMC chain. Another alternative but quite similar approach is the divide and conquer strategy presented by Rue (2001). While this approach shows better performance than the simple blocking strategies it has the disadvantage that the marginal density for one (possibly large) block needs to be calculated. The Lanczos algorithm, a Krylov subspace method (Saad, 2003) which has previously gained much attention for sampling from very large Gaussian densities (Aune et al., 2013; Simpson et al., 2013; Chow and Saad, 2014) offers an attractive alternative in these situations. It iteratively calculates an orthogonal projection $\tilde{\boldsymbol{x}}$ for the solution of $\boldsymbol{Q}^{1/2}\boldsymbol{x} = \boldsymbol{z}$ on a lower dimensional subspace. This produces a tridiagonal matrix and a set of vectors that depend on $\boldsymbol{Q}$ which can be used to find an approximation to $\boldsymbol{x}$. This procedure requires only matrix–vector products and is applicable to a wide range of precision matrices without the need to adjust blocking structures or visiting orders.

The rest of the paper is structured as follows. In Section 2, we briefly introduce the class of models we would like to address and give a basic MCMC sampling scheme that is widely used within this class of models for moderate data sizes. In Section 3, we show how the suggestions by Brezger and Lang (2006) can be combined with Krylov subspace methods in order to provide a general Bayesian framework for the estimation of large scale regression models. In addition, we present ways of how posterior samples for millions of parameters can be handled efficiently in the light of limited computational capacities. Section 4 examines the behavior and performance of the proposed method within an extensive simulation study. It is shown that our framework yields results for large-scale STAR models that are comparable to the ones obtained from state-of-the-art approaches fitting these types of models but requiring only low to moderate computational equipment. The data of these simulations have a similar structure as the data used in the applications in Section 5. Finally, we discuss and summarize our work in Section 6.