Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Q1 Extended dynamic generalized linear models: The two-parameter exponential family

Q2 M.A.O. Souza[a,*], H.S. Migon[b], J.B.M. Pereira[b]

[a] Fluminense Federal University, Brazil

[b] Federal University of Rio de Janeiro, Brazil

ARTICLE INFO

Article history:
Received 2 March 2015
Received in revised form 15 September 2016
Accepted 17 September 2016
Available online xxxx

Keywords:
Bayesian inference
Analytical approach
Conjugate updating

ABSTRACT

A Bayesian framework for estimation and prediction of dynamic models for observations from the two-parameter exponential family is developed. Different link functions are introduced to model both the mean and the precision in the exponential family allowing the introduction of covariates and time series components such as trend and seasonality. Conjugacy and analytical approximations are explored under the class of partially specified models to keep the computation fast. Due to the sequential nature of the proposed algorithm, all the advantages of sequential analysis, such as monitoring and intervention, can be applied to cope with the two-parameter exponential family models. The methodological novelties are illustrated with a simulation study and two applications to real data. The first application considers a well known financial time series regarding IBM stock returns modeled as following a gamma distribution. The second considers macroeconomic variables of the United Kingdom modeled as beta distributed data.

1. Introduction

Generalized linear models (GLM) are a standard class of models in data analysts' toolbox. Proposed by Nelder and Wedderburn (1972), GLM are widely used in many areas of knowledge. They allow modeling many different types of data via probabilistic description as an element of the exponential family relating the response mean and the linear predictor in nonlinear form. The GLM class is a useful alternative for data analysis since it accommodates skewness and heteroskedasticity, besides allowing analysis using the data in their original scale. The evolution of these models as well as details regarding inference, fitting, model checking, etc., are documented in the seminal book of McCullagh and Nelder (1989) and many other works in the recent literature.

The main criticism of the use of the one-parameter exponential family in certain applications is that samples are often too heterogeneous to be explained by a one-parameter family of models in the sense that the implicit mean–variance relationship in such a family is not supported by the data. To overcome this limitation, Gelfand and Dalal (1990) and Dey et al. (1997) introduced the two-parameter exponential family of models, which include the ones presented by Efron (1986) and Lindsay (1986) as special cases. They argue that the introduction of a second parameter allows taking into account the overdispersion usually present in the data, an issue that had been recognized by data analysts for many years.

During the 1990s, special attention was devoted to modeling the mean and the variance simultaneously. Taguchi type methods led to some efforts to jointly model the mean and the dispersion from designed experiments, avoiding data

* Correspondence to: Department of Statistics, Federal University of Rio de Janeiro, Rio de Janeiro, 21.941-909, Brazil.
E-mail addresses: mariana@est.uff.br (M.A.O. Souza), migon@im.ufrj.br (H.S. Migon), joao@dme.ufrj.br (J.B.M. Pereira).

http://dx.doi.org/10.1016/j.csda.2016.09.008

transformation, which is usually necessary to satisfy the assumptions of traditional linear models as in Nelder and Lee (2001). The process of quality improvement aims to minimize the product variation caused by different types of noise. Quality improvement must be implemented in the design stage via experiments to assess the sensitivity of different control factors that affect the variability and the mean of the process. Nelder and Lee (2001) discussed how the main ideas of a GLM can be extended to analyze Taguchi's experiments. From a static point of view, the Bayesian inference for this class of models is fully discussed in the papers previously cited, while some alternative aspects of MCMC are discussed in Cepeda and Gamerman (2005) and Cepeda et al. (2011).

Our aim is to extend the class of models introduced by Gelfand and Dalal (1990) and Dey et al. (1997) to deal with time series data and to propose a fast sequential algorithm for estimation and prediction in this class of models. To this end, an algorithm based on analytical approximations as an extension of the conjugate updating method proposed in West et al. (1985) is developed.

The remainder of the manuscript is organized as follows. Section 2 introduces the class of models of interest. In Section 3 the conjugate updating of West et al. (1985) is extended to the two-parameter exponential family. Section 4 illustrates the proposed method with a simulation study and two case studies: the first one models IBM stock returns, as described by Tsay (2002) and Triantafyllopoulos (2013), and the second one models data on the UK economy as beta distributed data. Section 5 concludes with a discussion and possible future research avenues.

2. Extended dynamic generalized linear models

In this section, the class of extended dynamic generalized linear models (EDGLM) is introduced. First a brief review of the two-parameter exponential family and dynamic generalized linear models is presented, mainly aiming to fix the notation used in the paper. A special parameterization of the two-parameter exponential family is presented in this section, which is useful to deal with data analysis when heterogeneity in the sample is greater than that explained by the variance function in the one-parameter exponential family. Distributions in this family are widely used in many applications in the current literature to deal with topics besides extra variability.

The two-parameter exponential family has the form

$$p(y|\theta,\phi) = a(y)\exp\left\{\phi[\theta d_1(y) + d_2(y)] - \rho(\theta,\phi)\right\}, \tag{1}$$

$y \in \Upsilon \subset \mathbb{R}$, where $a(\cdot)$ is a non-negative function, $d_1(\cdot)$ and $d_2(\cdot)$ are known real functions, $(\theta,\phi) \in \boldsymbol{\Theta} \times \boldsymbol{\Phi} \subseteq \mathbb{R} \times \mathbb{R}^+$ and $\exp\{\rho(\theta,\phi)\} = \int a(y)\exp\{\phi[\theta d_1(y) + d_2(y)]\}\,dy < \infty$. This is a suitable reparameterization of the general two-parameter exponential family as defined in Bernardo and Smith (1994).

This class includes many continuous distributions, such as the normal with unknown mean and variance, the inverse Gaussian and the beta distributions, parameterized by their mean and precision factors. The expression for the variances, as will be seen in Section 3.3, reveals the relevance of the precision parameter, ϕ, to control the model variance. Large values of ϕ correspond to more precise data or equivalently to data with smaller variance. Some discrete distributions are also included in this class, such as the binomial (with the sample size known) and Poisson distributions, taking the scale parameter as fixed and equal to one.

Among other interesting features of this class of distributions is the existence of a joint prior distribution for the parameters (θ,ϕ) in the form

$$p(\theta,\phi|\boldsymbol{\tau}) = \kappa(\boldsymbol{\tau})\exp\left\{\phi[\theta\tau_1 + \tau_2] - \tau_0\rho(\theta,\phi)\right\},$$

where $\boldsymbol{\tau} = (\tau_0,\tau_1,\tau_2)'$ and

$$\kappa(\boldsymbol{\tau})^{-1} = \iint \exp\left\{\phi[\theta\tau_1 + \tau_2] - \tau_0\rho(\theta,\phi)\right\} d\theta d\phi.$$

Let $\boldsymbol{\psi} = (\theta,\phi) \in \boldsymbol{\Psi} = \boldsymbol{\Theta} \times \boldsymbol{\Phi}$, to make the notation easier. Its prior mode and observed curvature matrix can be straightforwardly obtained by differentiating the expression above with respect to the parameters vector $\boldsymbol{\psi}$. More specifically, the mode and curvature matrix satisfy the equations $\tilde{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} \frac{\partial}{\partial\boldsymbol{\psi}}\log(p(\boldsymbol{\psi}|\boldsymbol{\tau}))$ and $J(\boldsymbol{\psi}) = -\frac{\partial^2}{\partial\boldsymbol{\psi}'\partial\boldsymbol{\psi}}\log(p(\boldsymbol{\psi}|\boldsymbol{\tau}))$. Then it follows, after some algebra, that

$$\begin{pmatrix} \phi\tau_1 - \tau_0\dfrac{\partial}{\partial\theta}\rho(\tilde{\boldsymbol{\psi}}) \\ \theta\tau_1 + \tau_2 - \tau_0\dfrac{\partial}{\partial\phi}\rho(\tilde{\boldsymbol{\psi}}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{and}$$

$$J(\boldsymbol{\psi}) = \begin{bmatrix} -\tau_{0t}\dfrac{\partial^2}{\partial\theta^2}\rho(\boldsymbol{\psi}) & \tau_1 - \tau_0\dfrac{\partial^2}{\partial\theta\partial\phi}\rho(\boldsymbol{\psi}) \\ \tau_1 - \tau_0\dfrac{\partial^2}{\partial\theta\partial\phi}\rho(\boldsymbol{\psi}) & -\tau_0\dfrac{\partial^2}{\partial\phi^2}\rho(\boldsymbol{\psi}) \end{bmatrix}.$$

Please cite this article in press as: Souza, M.A.O., et al., Extended dynamic generalized linear models: The two-parameter exponential family. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.09.008