Computational Statistics and Data Analysis 68 (2013) 375–387

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Mean field variational Bayesian inference for nonparametric regression with measurement error

Tung H. Pham[a], John T. Ormerod[b,*], M.P. Wand[c]

[a] *Institut de Mathématiques, École Polytechnique Fédérale de Lausanne, Station 8, CH-1015, Lausanne, Switzerland*
[b] *School of Mathematics and Statistics, University of Sydney, Sydney 2006, Australia*
[c] *School of Mathematical Sciences, University of Technology, Sydney, Broadway 2007, Australia*

ARTICLE INFO

Article history:
Received 18 January 2013
Received in revised form 8 July 2013
Accepted 8 July 2013
Available online 15 July 2013

Keywords:
Penalized splines
Classical measurement error
Markov chain Monte Carlo
Variational approximations

ABSTRACT

A fast mean field variational Bayes (MFVB) approach to nonparametric regression when the predictors are subject to classical measurement error is investigated. It is shown that the use of such technology to the measurement error setting achieves reasonable accuracy. In tandem with the methodological development, a customized Markov chain Monte Carlo method is developed to facilitate the evaluation of accuracy of the MFVB method.

1. Introduction

Flexible regression where the predictors are subject to measurement error continues to be an active area of research in the 2000s (Mallick et al., 2002; Liang et al., 2003; Carroll et al., 2004; Ganguli et al., 2005; Carroll et al., 2008) and is likely to be so in the 2010s. Carroll et al. (2006) offers a recent and comprehensive summary of the area.

Fitting and inference in such models is notoriously challenging. Berry et al. (2002) devised an elegant hierarchical Bayes approach to the simplest version of the problem and described Markov chain Monte Carlo (MCMC) based inference. Extensions have been considered by Carroll et al. (2004) and Ganguli et al. (2005). However, inference based on MCMC can be very slow for such models and may take hours if using BUGS (Lunn et al., 2000).

In this paper we investigate a faster mean field variational Bayes (MFVB) alternative to the problem. For an introduction to such techniques, see Bishop (2006) and Ormerod and Wand (2010) or Wand et al. (2011). We show that the transference of such technology to the measurement error setting achieves reasonable accuracy while being hundreds of times faster than MCMC. MFVB approximations to nonparametric regression problems with measurement error in the predictors are challenging due to spline basis functions entering the approximate posterior densities of the unobserved predictor. A streamlined discretization of these approximate posterior densities on a grid across the domain of the predictor is utilized to achieve computational efficiency.

In tandem with the methodological development a customized MCMC is developed to facilitate the evaluation of accuracy of the MFVB method. Both MCMC and MFVB are straightforward for all components of a nonparametric regression measurement error model, with the exception of the unobserved predictors. Approximate sampling from the

* Corresponding author. Tel.: +61 2 9351 5883; fax: +61 2 9351 4534.
E-mail addresses: tung.pham@epfl.ch (T.H. Pham), john.ormerod@sydney.edu.au, jtormerod@hotmail.com (J.T. Ormerod), Matt.Wand@uts.edu.au (M.P. Wand).

http://dx.doi.org/10.1016/j.csda.2013.07.014

full conditionals for the unobserved predictors can be performed efficiently using griddy-Gibbs sampling steps (Ritter and Tanner, 1992). Note that our MCMC and MFVB methods use an analogous approximation to the posterior distributions of the unobserved predictors.

After a brief introduction to MFVB methods (Section 2) we will develop these methods from the simplest case, simple linear regression (Section 3), and then extend these ideas to the more complex case of nonparametric regression with measurement error (Section 4) which could lay the foundation for more elaborate models such as additive models (see for example, Richardson and Green (2002) and Ganguli et al. (2005)). The methodology will be illustrated using a mix of simulated and real world examples (Section 5) and conclusions will be drawn (Section 6).

1.1. Notation

Throughout this paper *i.i.d.* is an abbreviation for *independent and identically distributed*. The notation $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means that $\mathbf{x}$ has a Multivariate Normal density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. x has an Inverse Gamma distribution, denoted by $x \sim \text{IG}(A, B)$, if and only if it has density $p(x) = B^A \Gamma(A)^{-1} x^{-A-1} \exp(-B/x)$, $x, A, B > 0$.

2. Elements of mean field variational Bayes

Let $\mathbf{D}$ be a vector of observed data, and $\boldsymbol{\theta}$ be a parameter vector with joint distribution $p(\mathbf{D}, \boldsymbol{\theta})$. In the Bayesian inferential paradigm decisions are made based on the posterior distribution $p(\boldsymbol{\theta}|\mathbf{D}) \equiv p(\mathbf{D}, \boldsymbol{\theta})/p(\mathbf{D})$ where $p(\mathbf{D}) \equiv \int p(\mathbf{D}, \boldsymbol{\theta}) d\boldsymbol{\theta}$. Let $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ be a partition of the parameter vector $\boldsymbol{\theta}$. Then mean field variational Bayes approximates $p(\boldsymbol{\theta}|\mathbf{D})$ by $q(\boldsymbol{\theta}) = \prod_{j=1}^M q(\boldsymbol{\theta}_j)$. It can be shown (see for example, Bishop (2006) and Ormerod and Wand (2010)) that the $q(\boldsymbol{\theta}_j)$s, often called q-densities, which minimize the Kullback–Leibler distance between $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\mathbf{D})$ defined by

$$\text{KL}(q, p) = \int q(\boldsymbol{\theta}) \log \left\{ \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{D})} \right\} d\boldsymbol{\theta} \tag{1}$$

are given by

$$q^*(\boldsymbol{\theta}_j) \propto \exp\left[E_{-q(\boldsymbol{\theta}_j)} \left\{p(\boldsymbol{\theta}_j|\text{rest})\right\}\right], \quad 1 \le j \le M, \tag{2}$$

where $E_{-q(\boldsymbol{\theta}_j)}$ denotes expectation with respect to $\prod_{k \ne j} q(\boldsymbol{\theta}_k)$. Note that only when (2) holds for each $q^*(\boldsymbol{\theta}_j)$, $1 \le j \le M$, is optimality obtained. Furthermore, a lower bound on the marginal log-likelihood is given by

$$\log p(\mathbf{D}) \ge \log \underline{p}(\mathbf{D}; q) = \int q(\boldsymbol{\theta}) \log \left\{ \frac{p(\mathbf{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta}. \tag{3}$$

It can be shown that the calculation of $q^*(\boldsymbol{\theta}_j)$ in (2) for fixed $\{q^*(\boldsymbol{\theta}_k)\}_{k \ne j}$ guarantees a monotonic increase in the lower bound (3) or equivalently a monotonic decrease in the Kullback–Leibler distance (1). Thus, an at least locally optimal $\{q(\boldsymbol{\theta}_j)\}_{1 \le j \le M}$ can be found by updating the $q^*(\boldsymbol{\theta}_j)$ in (2) sequentially until the lower bound (3) is judged to cease increasing.

To avoid notational clutter for a generic random variable v and density function $q(v)$ let

$$\mu_{q(v)} \equiv E_q(v) \quad \text{and} \quad \sigma_v^2 \equiv \text{Var}_q(v).$$

Also, in the special case that $q(v)$ is an Inverse Gamma density function we let $A_{q(v)}$ and $B_{q(v)}$ be the shape and rate parameters of $q(v)$ respectively, i.e. $v \sim \text{IG}(A_{q(v)}, B_{q(v)})$. Note $\mu_{q(1/v)} = A_{q(v)}/B_{q(v)}$. For a generic random vector $\mathbf{v}$ and density function $q(\mathbf{v})$ let

$$\boldsymbol{\mu}_{q(\mathbf{v})} \equiv E_q(\mathbf{v}) \quad \text{and} \quad \boldsymbol{\Sigma}_{q(\mathbf{v})} \equiv \text{Cov}_q(\mathbf{v}) = \text{covariance matrix of } \mathbf{v} \text{ under density } q(\mathbf{v}).$$

3. Simple linear regression with measurement error

We start with the simplest example of a measurement error model, where we want to perform a simple linear regression and the predictor is observed with error. Let

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad 1 \le i \le n, \tag{4}$$

where ε_i are i.i.d. $N(0, \sigma_\varepsilon^2)$. Here the responses, the y_is, are observed, but instead of observing $x_i \sim N(\mu_x, \sigma_x^2)$ we observe a corrupted version of x_i, w_i such that $w_i = x_i + v_i$, where v_i are i.i.d. $N(0, \sigma_v^2)$ random variables with σ_v^2 known.

For convenience we use independent priors with

$$\beta_0, \beta_1 \overset{\text{ind.}}{\sim} N(0, \sigma_\beta{}^2), \qquad \mu_x \sim N(0, \sigma_{\mu_x}^2), \qquad \sigma_x^2 \sim \text{IG}(A_x, B_x), \qquad \sigma_\varepsilon^2 \sim \text{IG}(A_\varepsilon, B_\varepsilon),$$

where $\sigma_\beta^2, A_\varepsilon, B_\varepsilon, A_x, B_x$ and $\sigma_{\mu_x}^2$ are positive hyperparameters.