Contents lists available at ScienceDirect

Neurocomputing

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

Bayesian variable selection for Gaussian process regression: Application to chemometric calibration of spectrometers

Tao Chen [a,*], Bo Wang [b]

[a] School of Chemical and Biomedical Engineering, Nanyang Technological University, Singapore 637459, Singapore
[b] Department of Mathematics, University of York, York YO10 5DD, UK

ARTICLE INFO

Article history:
Received 10 January 2010
Received in revised form
16 April 2010
Accepted 26 April 2010
Communicated by J. Zhang
Available online 5 May 2010

Keywords:
Gaussian process regression
Markov chain Monte Carlo
Model selection
Multivariate calibration
Variable selection

ABSTRACT

Gaussian processes have received significant interest for statistical data analysis as a result of the good predictive performance and attractive analytical properties. When developing a Gaussian process regression model with a large number of covariates, the selection of the most informative variables is desired in terms of improved interpretability and prediction accuracy. This paper proposes a Bayesian method, implemented through the Markov chain Monte Carlo sampling, for variable selection. The methodology presented here is applied to the chemometric calibration of near infrared spectrometers, and enhanced predictive performance and model interpretation are achieved when compared with benchmark regression method of partial least squares.

1. Introduction

Regression techniques are important data analysis tools in many scientific and engineering disciplines where empirical models are utilized to establish the relationship between two sets of variables. This paper is mainly concerned with the specific area of *chemometrics*, within which a major task is to calibrate various spectrometers such that the analyte properties (e.g. concentration) can be indirectly inferred from the measured spectra at hundreds or thousands of wavenumbers (or wavelengths) [1]. Fig. 1 gives the near infrared spectra of 120 samples of pharmaceutical tablets, where the objective is to predict the concentration of active substance from the absorbance recorded at 404 wavenumbers (i.e. 404 covariates). In this context, the conventional multiple linear regression is not applicable because of the high correlation between covariates, and the most widely used methods are principal component regression (PCR), partial least squares (PLS) and ridge regression [2,3]. The wide acceptance of these linear methods is largely due to the inherent linearity between the analyte properties and spectral absorbance as stated in the *Beer–Lambert's law* [4]. However, nonlinear calibration methods have also been proposed for practical problems where the linearity does not hold, including neural networks [5], support vector machine and its variants [6,7].

Recently, there has been significant interest in the Gaussian process regression model. Initially proposed by [8], Gaussian process was viewed as an alternative approach to neural networks since it was demonstrated that a large class of Bayesian regression models, based on neural networks, converged to a Gaussian process in the limit of an infinite network [9]. Gaussian processes can also be derived from the perspective of non-parametric Bayesian regression [10], by directly placing Gaussian prior distributions over the space of regression functions. Empirical comparative studies have confirmed the outstanding performance of Gaussian process regression with respect to other non-linear models [11–13]. As a result, Gaussian process models have been widely applied to various problems in statistics and engineering [11,13–18].

In developing a Gaussian process model for regression on a large number of covariates, it is often desirable to reduce the dimension of the variables to alleviate the computational burden and to improve the prediction accuracy. This is typically realized by either projecting the original covariates onto lower-dimensional space, for example using principal component analysis (PCA) [11], or selecting a subset of these covariates. Compared with the projection techniques, variable selection attains the additional advantage of improved interpretability as to which covariates are the most informative to prediction. Therefore, the task of variable selection with respect to Gaussian process regression is considered in this paper. This is a special case of model selection problems, since a certain set of selected variables corresponds to a particular model.

[*] Corresponding author. Tel.: +65 6513 8267; fax: +65 6794 7553.
E-mail addresses: chentao@ntu.edu.sg (T. Chen), bw527@york.ac.uk (B. Wang).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.04.014

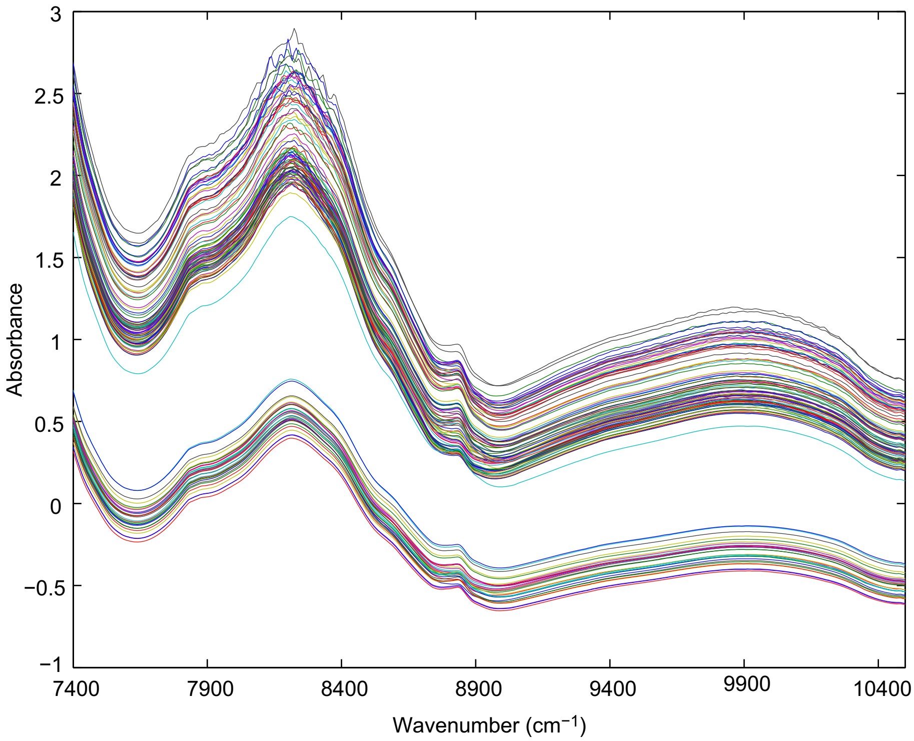

Fig. 1. Near infrared spectra of 120 tablet samples.

In the field of chemometric calibration modelling, variable selection strategies are typically based on a PLS regression model whilst optimizing predictive performance by selecting/removing the covariates. For example, iterative PLS [19] starts with the random selection of a small number of variables, with variables being added or removed based on the cross-validation error. An alternative approach is uninformative variable elimination based on an analysis of the PLS regression coefficients [20]. A third method widely reported in the literature is that of genetic algorithms (GAs) that were originally proposed as a family of stochastic optimization approaches that mimic the principles of genetics and natural selection [21]. GAs have been successfully applied for variable selection in spectroscopic calibration [22,23].

More recently Bayesian approach to variable selection has received considerable attention, where the main focus has been on linear regression using Markov chain Monte Carlo (MCMC) methods. Initially the Gibbs sampling algorithm was proposed to sample the posterior of the indicators [24,25]. This was followed by its extension to multiple responses [26], and its implementation using the Metropolis–Hastings sampling algorithm [27]. Dellaportas et al. [28] reviewed and compared several MCMC methods for variable selection. Basis function based regression models, for example splines [29–31] and wavelets [32], were proposed with the selection of the basis function being based on similar methodologies to those of variable selection.

The challenge with variable selection in a Gaussian process is that the posterior distribution of the "hyper-parameters" (to be defined subsequently) is not analytically available and needs to be sampled. As a consequence the tasks of model selection and hyper-parameter estimation are coupled, and this issue is addressed in this paper by sampling the model and hyper-parameters alternately. The rest of the paper is organized as follows. A brief description of Gaussian process regression is provided in Section 2. An overview of the Bayesian model selection techniques is described in Section 3, where the discussion focuses on MCMC based approaches. Section 4 presents the MCMC method for variable selection, and the proposed method is demonstrated through application study in Section 5. Finally conclusions are drawn in Section 6.

2. Gaussian process regression

Consider the case where a training set of N observations is available, $\{\mathbf{x}_i, y_i;\ i=1,\ldots,N\}=\{\mathbf{X},\mathbf{y}\}$, where $\mathbf{x}_i$ is a vector of p covariates, and y_i is the scalar response. A Gaussian process prior for regression is then defined in such a way that the regression function has a Gaussian process prior with zero mean and covariance function $C(y_i,y_j)=C(\mathbf{x}_i,\mathbf{x}_j;\boldsymbol{\theta})$. The covariance is a function of covariates given the hyper-parameters $\boldsymbol{\theta}$. An example of such a covariance function is

$$C(\mathbf{x}_i,\mathbf{x}_j;\boldsymbol{\theta})=a_0+a_1\sum_{d=1}^{p}x_{id}x_{jd}+v_0\exp\left(-w\sum_{d=1}^{p}(x_{id}-x_{jd})^2\right)+\delta_{ij}\sigma^2 \tag{1}$$

where $\mathbf{x}_i=(x_{i1},\ldots,x_{ip})$; $\delta_{ij}=1$ if $i=j$, otherwise it is equal to zero. We denote $\boldsymbol{\theta}=(a_0,a_1,v_0,w,\sigma^2)$ "hyper-parameters" to differentiate Gaussian processes from parametric regression. In this covariance function, the first two terms represent constant bias and linear correlation, respectively. The exponential term is similar to the form of a radial basis function, and defines the correlation between the responses and nearby covariates. Finally σ^2 captures random noise. By combining linear and non-linear terms in the covariance function, the Gaussian process is capable of handling both linear and non-linear data structures [11]. Other forms of the covariance function are discussed in [10,33].

The training of a Gaussian process, i.e. the inference of $\boldsymbol{\theta}$, can be carried out using maximum likelihood estimation, or using maximum a posterior estimation if prior distributions are given for the hyper-parameters. Since the posterior distribution of $\boldsymbol{\theta}$ is generally of complicated form, it is desirable to use MCMC