

2008 Special Issue

Variational Bayesian least squares: An application to brain–machine interface data[☆]

Jo-Anne Ting^{a,*}, Aaron D'Souza^b, Kenji Yamamoto^c, Toshinori Yoshioka^d, Donna Hoffman^e, Shinji Kakei^f, Lauren Sergio^g, John Kalaska^h, Mitsuo Kawato^d, Peter Strick^e, Stefan Schaal^{a,d}

^a University of Southern California, Los Angeles, CA, 90089, USA

^b Google, Inc., Mountain View, CA 94043, USA

^c National Institute of Radiological Sciences, Chiba 263-8555, Japan

^d ATR Computational Neuroscience Laboratories, Kyoto 619-0288, Japan

^e University of Pittsburgh, Pittsburgh, PA 15261, USA

^f Tokyo Metropolitan Institute for Neuroscience, Tokyo 183-8526, Japan

^g York University, Toronto, Ontario, Canada M3J 1P3

^h Université de Montréal, Montréal, Canada H3C 3J7

ARTICLE INFO

Article history:

Received 15 December 2007

Received in revised form

10 June 2008

Accepted 17 June 2008

Keywords:

High-dimensional regression

Variational Bayesian methods

Linear models

Dimensionality reduction

Feature selection

Brain–machine interfaces

EMG prediction

Statistical learning

ABSTRACT

An increasing number of projects in neuroscience require statistical analysis of high-dimensional data, as, for instance, in the prediction of behavior from neural firing or in the operation of artificial devices from brain recordings in brain–machine interfaces. Although prevalent, classical linear analysis techniques are often numerically fragile in high dimensions due to irrelevant, redundant, and noisy information. We developed a robust Bayesian linear regression algorithm that automatically detects relevant features and excludes irrelevant ones, all in a computationally efficient manner. In comparison with standard linear methods, the new Bayesian method regularizes against overfitting, is computationally efficient (unlike previously proposed variational linear regression methods, is suitable for data sets with large numbers of samples and a very high number of input dimensions) and is easy to use, thus demonstrating its potential as a drop-in replacement for other linear regression techniques. We evaluate our technique on synthetic data sets and on several neurophysiological data sets. For these neurophysiological data sets we address the question of whether EMG data collected from arm movements of monkeys can be faithfully reconstructed from neural activity in motor cortices. Results demonstrate the success of our newly developed method, in comparison with other approaches in the literature, and, from the neurophysiological point of view, confirms recent findings on the organization of the motor cortex. Finally, an incremental, real-time version of our algorithm demonstrates the suitability of our approach for real-time interfaces between brains and machines.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

In recent years, there has been growing interest in large scale

[☆] This research was supported in part by National Science Foundation grants ECS-0325383, IIS-0312802, IIS-0082995, ECS-0326095, ANI-0224419, a NASA grant AC#98-516, an AFOSR grant on Intelligent Control, the ERATO Kawato Dynamic Brain Project funded by the Japanese Science and Technology Agency, the ATR Computational Neuroscience Laboratories, funds from the Department of Veterans Affairs, Medical Research Service, NIH grant P01 NS044393 (PLS), and funds from CIHR and FRSQ. KY was partially supported by a Japan Society for the Promotion of Sciences Postdoctoral Fellowship for research abroad.

* Corresponding address: University of Southern California, Hedco Neurosciences Building, 3641 Watt Way, Los Angeles, CA 90089, USA. Tel.: +1 213 740 6717; fax: +1 213 740 5687.

E-mail address: joanneti@usc.edu (J.-A. Ting).

analyses of brain activity, with respect to associated behavioral variables. For instance, projects can be found in the area of brain–machine interfaces, where neural firing is directly used to control an artificial system like a robot (Chapin, Moxon, Markowitz, & Nicolelis, 1999; Hochberg et al., 2006; Lebedev & Nicolelis, 2006; Nicolelis, 2001; Nicolelis & Ribeiro, 2006; Taylor, Tillery, & Schwartz, 2002), or where non-invasive brain signals serve to either control a cursor on computer screen (Wolpaw & McFarland, 2004), or to classify visual stimuli presented to a subject (Haynes & Rees, 2005; Kamitani & Tong, 2004). In such scenarios, the brain signals to be processed are typically high dimensional, in the order of hundreds or thousands of inputs, with large numbers of redundant and irrelevant signals. Linear modeling techniques like linear regression are among

the primary analysis tools for such data (Lebedev & Nicolelis, 2006; Musallam, Corneil, Greger, Scherberger, & Andersen, 2004; Wessberg & Nicolelis, 2004). However, the computational problem of data analysis not only involves data fitting, but also requires that the model extracted from the data has good generalization properties. This issue is crucial for predicting behavior from future neural recordings, e.g., for continual on-line interpretation of brain activity to control prosthetic devices, or for longitudinal scientific studies of information processing in the brain. Surprisingly, robust linear modeling of high-dimensional data is non-trivial, as the danger of fitting noise and of encountering numerical problems is high. Classical techniques like ridge regression, stepwise regression, subset selection techniques, or Partial Least Squares regression (Wold, 1975) are known to be prone to overfitting, and may often require careful human supervision to ensure useful results. Other methods such as Least Absolute Shrinkage and Selection Operator (LASSO) regression (Tibshirani, 1996) attempt to shrink certain regression coefficients to zero, resulting in interpretable models that are sparse. However, LASSO regression has an open parameter that needs to be set, using either n -fold cross-validation or manual hand-tuning.

In this paper, we will focus on how to improve linear data analysis for the high-dimensional scenarios described above, with a view towards developing a “black box” approach that automatically detects the most relevant input dimensions for generalization and excludes other dimensions in a statistically sound way. We are particularly interested in situations where the data contains a very large quantity of samples and the number of input dimensions is very high, as in brain–machine interfaces. For this purpose, we investigate a full Bayesian treatment of linear regression with automatic relevance detection (Neal, 1994) that is computationally efficient and suitable for large amounts of very high-dimensional data. This algorithm can be formulated in closed form with the help of a variational Bayesian approximation, and it introduces “probabilistic backfitting” for linear regression, a key component which contributes greatly towards the algorithm’s computational efficiency. Besides several synthetic data evaluations, we apply the algorithm, named Variational Bayesian Least Squares (VBLS) (Ting et al., 2005), to the reconstruction of EMG data from motor cortical firing from data sets collected by Sergio and Kalaska (1998) and Kakei, Hoffman, and Strick (1999, 2001). This data analysis addresses important neurophysiological questions in terms of whether motor cortical neurons can directly predict EMG traces (Bennett & Lemon, 1996; McKiernan, Marcario, Karrer, & Cheney, 1998; Morrow & Miller, 2003; Todorov, 2000; Townsend, Paninski, & Lemon, 2006), whether motor cortices have a muscle-based topological organization, and whether information in motor cortices should be used to predict behavior in future brain–machine interfaces. Our main focus in this paper is on the statistical analysis of these kinds of data. Comparisons with classical linear analysis techniques and a brute force combinatorial model search (which was executed on a cluster computer) demonstrate that our VBLS algorithm indeed achieves the “black box” quality of a statistical analysis technique that requires no tuning of parameters by the user.

This paper describes in detail the VBLS algorithm and its application to the EMG reconstruction problem by building and extending our prior work in D’Souza, Vijayakumar, and Schaal (2004) and Ting et al. (2005). We discuss the neurophysiological implications of our analyses and present a real-time version of VBLS in order to simulate an application in real-time brain machine interfaces.

2. High dimensional regression

Before developing our VBLS algorithm, it is useful to briefly revisit classical linear regression techniques. Assuming there are N observed data samples in the data set $D = \{\mathbf{x}_i, y_i\}_{i=1}^N$ (where $\mathbf{x}_i \in \mathfrak{R}^{d \times 1}$ are inputs and y_i are scalar outputs), the standard model for linear regression is:

$$y_i = \sum_{m=1}^d b_m x_{im} + \epsilon \quad (1)$$

where \mathbf{b} is the regression vector made up of b_m components, d is the number of input dimensions, and ϵ is additive mean-zero noise. The Ordinary Least Squares (OLS) estimate of the regression vector is $\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, where $\mathbf{X} \in \mathfrak{R}^{N \times d}$ consists of vectors \mathbf{x}_i arranged in its rows and $\mathbf{y} \in \mathfrak{R}^{N \times 1}$ has coefficients y_i . The main problem with OLS regression in high-dimensional input spaces is that the full rank assumption of $(\mathbf{X}^T \mathbf{X})^{-1}$ is often violated due to underconstrained data sets. Ridge regression (Hoerl & Kennard, 1970) can “fix” such problems numerically by stabilizing the matrix inversion with a diagonal term $(\mathbf{X}^T \mathbf{X} + \alpha \mathbf{I})^{-1}$, but usually introduces uncontrolled bias. Additionally, if the input dimensionality exceeds around 1000 dimensions, the matrix inversion can become prohibitively computationally expensive.

Several ideas exist how to improve over OLS. First, stepwise regression (Draper & Smith, 1981) can be employed. However, stepwise regression has been strongly criticized for its potential for overfitting and its inconsistency in the presence of collinearity in the input data (Derksen & Keselman, 1992). To deal with such collinearity directly, dimensionality reduction techniques like Principal Components Regression (PCR) (Massey, 1965) are useful. These methods retain directions in an input space with large variance, regardless of whether the directions influence the prediction (Schaal, Vijayakumar, & Atkeson, 1998), and can even eliminate low variance inputs that may have high predictive power for the outputs (Frank & Friedman, 1993). Another class of linear regression methods is projection regression techniques, most notably Partial Least Squares (PLS) regression (Wold, 1975). PLS regression performs computationally inexpensive $O(d)$ univariate regressions along projection directions, chosen according to the correlation between inputs and outputs. While slightly heuristic in nature, PLS regression is a surprisingly successful algorithm for ill-conditioned and high-dimensional regression problems, although it also has a tendency towards overfitting (Schaal et al., 1998). There are also more efficient methods for matrix inversion (Hastie & Tibshirani, 1990; Strassen, 1969), but these methods assume a well-condition regression problem *a priori* and degrade in the presence of collinearities in inputs. Finally, there is a class of sparsity inducing methods such as LASSO regression (Tibshirani, 1996) that attempt to shrink certain regression coefficients in the solution to zero by using an L1 penalty norm (instead of an L2 penalty norm used by ridge regression). These methods are suitable for high-dimensional data sets, at the expense of requiring an open parameter (i.e., a fixed bound on the penalty norm) that needs to be set using cross-validation. Note that previous methods of sparse variational linear regression have been proposed by Bishop (2006) and Tipping (2001), however these are not computationally efficient and are unsuitable for large amounts of high-dimensional data.

We will use some of the previously described methods for comparison in the Evaluation section. In particular, we will compare our proposed algorithm to the following methods: (i) OLS regression, (ii) ridge regression with an empirically tuned ridge value, (iii) stepwise regression, (iv) PLS regression and (v) LASSO regression. In the next section, we will introduce

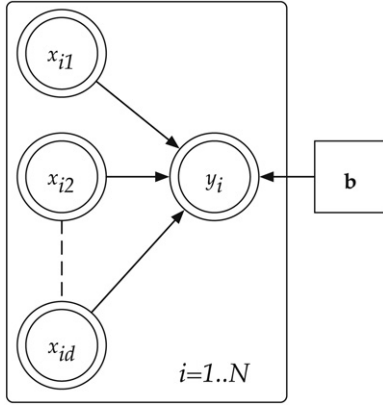


Fig. 1. Graphical model for linear regression. Random variables are in circular nodes, observed random variables are in double circles, and point estimated parameters are in square nodes.

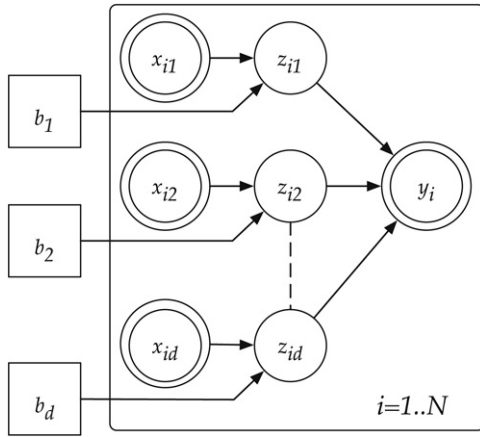


Fig. 2. Graphical model for Probabilistic Backfitting. Random variables are in circular nodes, observed random variables are in double circles, and point estimated parameters are in square nodes.

a linear regression algorithm in a Bayesian framework that *automatically* regularizes against problems of overfitting (in contrast, LASSO regression has an open parameter that requires cross-validation in order to find its optimal value). Additionally, the iterative nature of the algorithm – due to its formulation as an Expectation-Maximization problem (Dempster, Laird, & Rubin, 1977) – avoids the computational cost and numerical problems of matrix inversions that is faced in high-dimensional OLS regression and in Bishop (2006) and Tipping (2001). Thus, VBLS addresses the two major problems of high-dimensional OLS regression simultaneously. Note, however, that if accurate results are needed (and computational resources are unlimited) for data sets with fully relevant input dimensions, VBLS is not as efficient as the matrix inversion in OLS. The advantage of VBLS arises when dealing with high dimensional input spaces, serving as an efficient and robust “automatic” regression method. Conceptually, the algorithm can be interpreted as a Bayesian version of either backfitting or Partial Least Squares regression.

3. Variational Bayesian least squares

Figs. 1–3 illustrate the progression of graphical models that we need to develop a robust Bayesian version of linear regression. Fig. 1 depicts the standard linear regression model. Part of the inspiration for our algorithm comes from PLS regression, motivated by the question of how to find maximally predictive projections in input space, which is also part of various other

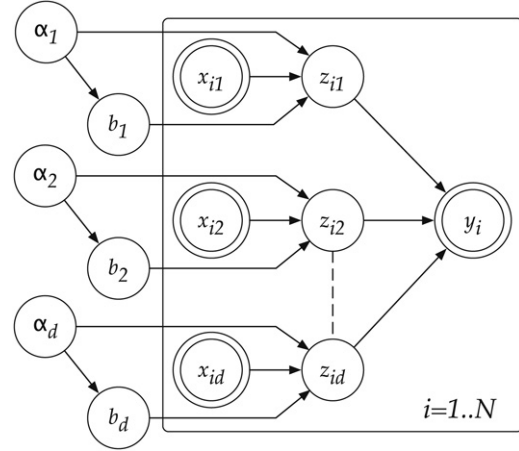


Fig. 3. Graphical model for Variational Bayesian Least Squares. Random variables are in circular nodes, observed random variables are in double circles, and point estimated parameters are in square nodes.

“subset” selection techniques in regression (Wessberg & Nicolelis, 2004). Indeed, if we knew the optimal projection direction of the input data, the entire regression problem could be solved by a univariate regression between the projected data and the outputs: this optimal projection direction is simply the true gradient between inputs and outputs. Since we do not know this projection direction, we now encode its coefficients as hidden variables, in the tradition of Expectation-Maximization (EM) algorithms (Dempster et al., 1977). Fig. 2 shows the corresponding graphical model. The unobservable variables z_{im} (where $i = 1, \dots, N$ denotes the index into the data set of N data points) are the result of the input variables being projected on the respective projection direction component (i.e., b_m). Then, the z_{im} ’s are summed up to form a predicted output y_i . More formally, we can modify the linear regression model in Eq. (1) to become:

$$y_i = \sum_{m=1}^d z_{im} + \epsilon_y \tag{2}$$

$$z_{im} = b_m x_{im} + \epsilon_{z_m}. \tag{3}$$

For a probabilistic treatment with EM, we make a standard normal assumption of all distributions in form of:

$$y_i | \mathbf{z}_i \sim \text{Normal}(\mathbf{1}^T \mathbf{z}_i, \psi_y) \tag{4}$$

$$z_{im} | x_{im} \sim \text{Normal}(b_m x_{im}, \psi_{z_m})$$

where $\mathbf{1} = [1, 1, \dots, 1]^T$. While this model is still identical to OLS, notice that in the graphical model of Fig. 2, the regression coefficients b_m are behind the fan-in to the outputs y_i . We call this model Probabilistic Backfitting, since the resulting derived update equation for the regression coefficient b_m can be viewed as a probabilistic version of backfitting. Given the data D , we can view this new regression model as an EM problem and maximize the incomplete log likelihood $\log p(\mathbf{y} | \mathbf{X})$ by maximizing the expected complete log likelihood ($\log p(\mathbf{y}, \mathbf{Z} | \mathbf{X})$), where:

$$\log p(\mathbf{y}, \mathbf{Z} | \mathbf{X}) = -\frac{N}{2} \log \psi_y - \frac{1}{2\psi_y} \sum_{i=1}^N (y_i - \mathbf{1}^T \mathbf{z}_i)^2 - \frac{N}{2} \sum_{m=1}^d \log \psi_{z_m} - \sum_{m=1}^d \frac{1}{2\psi_{z_m}} \sum_{i=1}^N (z_{im} - b_m x_{im})^2 + \text{const} \tag{5}$$

where $\mathbf{Z} \in \mathfrak{R}^{N \times d}$ consists of z_{im} components. The resulting EM updates require standard manipulations of normal distributions

