Journal of Multivariate Analysis 101 (2010) 1574–1593

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Statistical inference of minimum BD estimators and classifiers for varying-dimensional models

Chunming Zhang

Department of Statistics, University of Wisconsin, 53706 Madison, WI, United States

ARTICLE INFO

Article history:
Received 26 February 2009
Available online 4 March 2010

AMS 2000 subject classifications:
primary 62F12
62F05
secondary 62J02
60J12

Keywords:
A diverging number of parameters
Exponential family
Hemodynamic response function
Loss function
Optimal Bayes rule

ABSTRACT

Stochastic modeling for large-scale datasets usually involves a varying-dimensional model space. This paper investigates the asymptotic properties, when the number of parameters grows with the available sample size, of the minimum-BD estimators and classifiers under a broad and important class of Bregman divergence (BD), which encompasses nearly all of the commonly used loss functions in the regression analysis, classification procedures and machine learning literature. Unlike the maximum likelihood estimators which require the joint likelihood of observations, the minimum-BD estimators are useful for a range of models where the joint likelihood is unavailable or incomplete. Statistical inference tools developed for the class of large dimensional minimum-BD estimators and related classifiers are evaluated via simulation studies, and are illustrated by analysis of a real dataset.
© 2010 Elsevier Inc. All rights reserved.

1. Introduction

In many fields of applications, the dimension (number) p of model space (parameters) depends on the sample size n. Examples include the X-ray crystallography [1,2] and the autoregressive models in times series. In the literature, Drost [3] developed the goodness-of-fit tests for location-scale models when the number of classes tends to infinity; Murphy [4] developed testing for a time dependent coefficient in Cox's regression model. This paper is motivated from issues in two important and challenging applications.

1.1. fMRI time series: a diverging number of parameters

Functional magnetic resonance imaging (fMRI) is a recent and exciting method that allows investigators to determine which areas of the brain are involved in a cognitive task. Following Ward [5] and Worsley et al. [6], a single-voxel fMRI time-series $\{s(t_i), y(t_i)\}_{i=1}^n$, for a given scan and a given subject, can be captured by the convolution model

$$y(t) = d(t) + s * h(t) + \varepsilon(t), \quad t = t_1, \dots, t_n, \tag{1.1}$$

where $*$ denotes the convolution operator, $y(t)$ is the measured noisy fMRI signal, $s(t)$ is the external input stimulus (which could be from a design either block- or event-related and where $s(t) = 1$ or 0 indicates the presence or absence of a stimulus), $h(t)$ is the hemodynamic response function (HRF) at time t after neural activity, $d(t)$ is a slowly drifting baseline, and the errors $\varepsilon(t_i)$ are zero-mean and temporally autocorrelated. Similar models can be found in [7]. Refer to [8] and references therein for a recent review of statistical issues and methods in fMRI data analysis.

E-mail address: cmzhang@stat.wisc.edu.

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.02.009

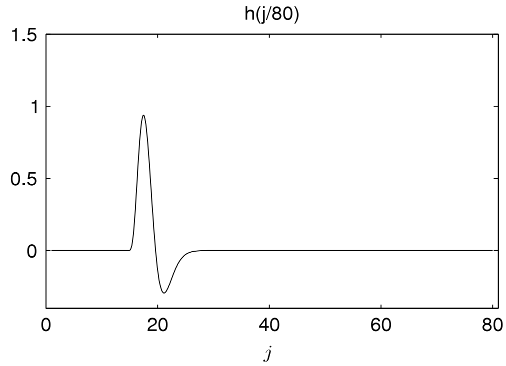

Fig. 1. An illustrative plot of HRF $h(t_j)$ with $n = 80$.

Of primary interest to neuroscientists is the estimation and hypothesis testing of the underlying HRF. Typically, the peak value of the HRF $h(\cdot)$ is reached after a short delay of the stimulus and drops quickly to zero. A typical example of $h(\cdot)$, given in [9], is plotted in Fig. 1. Clearly, the region $\{t : h(t) \neq 0\}$ is sparse in its temporal domain. Thus, to obtain statistically more efficient estimates of the HRF associated with event-related fMRI experiments, the sparsity of the HRF needs to be taken into account. We thus suppose that $h(t) = 0$ for $t > t_{p_n}$ and focus on estimating the first p_n values of $h(t_i)$, where p_n is less than n, the length of the fMRI time series. In neuroimaging studies, the temporal drift $d(\cdot)$ is a nuisance function and usually approximated by a (at most third order) polynomial; see for example, the popular imaging analysis tool AFNI at http://afni.nimh.nih.gov/afni/ [10,6]. As such, (1.1) is re-expressed as

$$\mathbf{y} = \widetilde{\mathbf{T}}\widetilde{\boldsymbol{\alpha}} + \mathbf{S}\mathbf{h} + \boldsymbol{\epsilon}, \tag{1.2}$$

where $\mathbf{y} = (y(t_1), \dots, y(t_n))^T$,

$$\widetilde{\mathbf{T}} = \begin{bmatrix} 1 & t_1 & t_1^2 & t_1^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 & t_n^3 \end{bmatrix}, \qquad \mathbf{S} = \begin{bmatrix} s(0) & 0 & \cdots & 0 \\ s(t_2 - t_1) & s(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ s(t_{p_n} - t_1) & s(t_{p_n} - t_2) & \cdots & s(0) \\ \vdots & \vdots & \cdots & \vdots \\ s(t_n - t_1) & s(t_n - t_2) & \cdots & s(t_n - t_{p_n}) \end{bmatrix}$$

is the $n \times p_n$ Toeplitz matrix, $\boldsymbol{\epsilon} = (\epsilon(t_1), \dots, \epsilon(t_n))^T$, and $\widetilde{\boldsymbol{\alpha}} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)^T$ and $\mathbf{h} = (h(t_1), \dots, h(t_{p_n}))^T$ are both vectors of unknown parameters.

Clearly, p_n here grows with n but is excessively below n. For the regression problem (1.2), the large-dimensional vector $\mathbf{h}$ can be estimated via weighted least-squares using the quadratic loss. In practice, however, p_n, termed the "intrinsic dimensionality of fMRI data" [11], is unknown for real fMRI data. Indeed, as far as we know, all published work for fMRI assumes that $p_n = p$ is a known fixed number, followed by traditional parametric inference based on asymptotic derivations of fixed-dimensional estimators. To reduce modeling biases due partly to the fixed choice of p, statistical inference based on asymptotic results which allow the dimension p_n to depend on n is desired. This motivates us to consider a more realistic relation between p_n and n,

$$\text{dimension} : p_n \text{ varies with } n \text{ or even } p_n \to \infty \text{ at a certain rate as } n \to \infty. \tag{1.3}$$

1.2. Statistical learning: regression and classification under general loss

In statistical learning, the primary goals of regression and classification seem to be kept separate. Regression methods concern the "orderable" output variable and aim to estimate the regression function at points of the input variable, whereas the primary interest of classification rules for the "categorical" output variable is to forecast the most likely class label for the output.

As discussed in [12], both regression and classification can be viewed from the common perspective of real valued prediction. Namely, the goal of a supervised learning algorithm is to use the training samples to construct a prediction rule for a future output at the observed value of the input variable. Depending on the nature of the output variable, the predictive error is quantified by different error measures. For example, the quadratic loss function, as utilized in the previous brain fMRI data, has nice analytical properties and is usually used in regression analysis. However, the quadratic loss is not always adequate in classification problems where the misclassification loss, deviance loss (or the negative log-likelihood) and exponential loss are more realistic and commonly used in classification.

Download English Version:

https://daneshyari.com/en/article/1146548

Download Persian Version:

https://daneshyari.com/article/1146548

Daneshyari.com