Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2007) 406–422
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Input selection and shrinkage in multiresponse linear regression

Timo Similä*, Jarkko Tikka

Laboratory of Computer and Information Science, Helsinki University of Technology, P.O. Box 5400, FI-02015 HUT, Helsinki, Finland

Available online 17 February 2007

Abstract

The regression problem of modeling several response variables using the same set of input variables is considered. The model is linearly parameterized and the parameters are estimated by minimizing the error sum of squares subject to a sparsity constraint. The constraint has the effect of eliminating useless inputs and constraining the parameters of the remaining inputs in the model. Two algorithms for solving the resulting convex cone programming problem are proposed. The first algorithm gives a pointwise solution, while the second one computes the entire path of solutions as a function of the constraint parameter. Based on experiments with real data sets, the proposed method has a similar performance to existing methods. In simulation experiments, the proposed method is competitive both in terms of prediction accuracy and correctness of input selection. The advantages become more apparent when many correlated inputs are available for model construction.

Keywords: Subset selection; Variable selection; Constrained regression; Multivariate regression; Convex optimization; Cone programming

1. Introduction

Multiresponse regression is the task of estimating several response variables using a common set of input variables. There are two approaches to the problem. Either a separate model is built for each response variable, or a single model is used to estimate all the responses simultaneously. Breiman and Friedman (1997) and Srivastava and Solanky (2003) present simultaneous estimation techniques that have advantages over the separate model building, especially when the responses are correlated. Correlation among the responses is typical in many applications, for instance, in the field of chemometrics (Burnham et al., 1999). In this article, the focus is on linear simultaneous models.

Many input variables are usually available for model construction. However, some of the inputs may be weakly correlated with the responses and some others may be redundant in that they are highly correlated with the other inputs. A small number of observations compared to the number of inputs causes the problem of *overfitting*: the model fits well on training data but generalizes poorly. Highly correlated inputs cause the problem of *collinearity*: model interpretation is misleading as the importance of an input in the model can be compensated by another input. Traditional methods for meeting these problems are pure input selection (Sparks et al., 1985), regularization or shrinking (Breiman and Friedman, 1997; Srivastava and Solanky, 2003), and subspace methods (Abraham and Merola, 2005). Shrinking means that the regression coefficients are constrained such that the unimportant inputs tend to have coefficient values close

* Corresponding author. Tel.: +358 9 4513360; fax: +358 9 4513277.
E-mail addresses: timo.simila@hut.fi (T. Similä), tikka@cis.hut.fi (J. Tikka).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.01.025

to zero. In the subspace approach the data are projected onto a smaller subspace in which the model is fitted. Input selection differs from the two other techniques as some of the inputs are completely left out of the model.

Practical benefits of input selection include aid in model interpretation, economic efficiency if measured inputs have costs, and computational efficiency due to simplicity of the model. Commonly used criteria for input selection are tests of statistical significance, information criteria, and prediction error (Bedrick and Tsai, 1994; Barrett and Gray, 1994; Sparks et al., 1985). These criteria only rank combinations of inputs and some greedy stepwise method is typically applied to find promising combinations. However, the greedy stepwise methods may fail to recognize important combinations of inputs, especially when the inputs are highly correlated (Derksen and Keselman, 1992). Better results can be obtained by incorporating shrinking in the selection strategy (Breiman, 1996; Similä and Tikka, 2006). Bayesian methods offer another approach (Brown et al., 2002), which is theoretically sound but may be a bit technical from a practical point of view. Recently, more straightforward methods have emerged in the statistical and signal processing communities, apparently through independent research efforts (Turlach et al., 2005; Cotter et al., 2005; Malioutov et al., 2005; Tropp, 2006). These methods either constrain or penalize the model fitting in a way that input selection and shrinking occur simultaneously. As a common denominator, the estimation is formulated as a single convex optimization problem. From now on, the family of this type of methods is called as simultaneous variable selection (SVS).

We consider a SVS method, which is used in the signal processing community (Cotter et al., 2005; Malioutov et al., 2005). The importance of an input in the model is measured by the 2-norm of the regression coefficients associated with the input, and that is why the method is denoted by L_2-SVS. The error sum of squares is minimized while constraining the sum of the importances over all the input variables. We also discuss a variant of SVS, where the ∞-norm is used instead of the 2-norm. L_∞-SVS is proposed by Turlach et al. (2005) in the statistical and Tropp (2006) in the signal processing community. The main contributions of this article are a formal analysis of the L_2-SVS problem and a numerical solver, which takes advantage of the structure of the problem. Furthermore, we present an efficient algorithm for following the path of solutions as a function of the constraint parameter. The existing SVS articles do not consider the solution path, although it is highly useful in practical problems, where the constraint parameter must be fixed by cross-validation or related techniques.

The rest of this article is organized as follows. In Section 2, we introduce the L_2-SVS estimate and position it with respect to related research. In Section 3, we derive the optimality conditions and propose algorithms for solving the L_2-SVS problem. Two types of comparisons are presented in Section 4. Firstly, several real world data sets are analyzed. Secondly, simulation experiments are carried out to explore the effect of collinearity among the input variables. Section 5 concludes the article.

2. L_2-SVS estimate and discussion of related work

Suppose that we have q response variables and m input variables from which we have n observations. The response data are denoted by an $n \times q$ matrix $\boldsymbol{Y}$ and the input data by an $n \times m$ matrix $\boldsymbol{X}$. All the variables are assumed to have zero means and similar scales. We focus on a linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^* + \boldsymbol{E}, \tag{1}$$

where $\boldsymbol{W}^*$ is an $m \times q$ matrix of regression coefficients and $\boldsymbol{E}$ is an $n \times q$ matrix of noise. Some rows of the matrix $\boldsymbol{W}^*$ are assumed to have only zero elements, which means that the corresponding input variables do not contribute to the response variables at all. In order to simplify the following notation, we further define

$$\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_m] = [\underline{\boldsymbol{x}}_1 \cdots \underline{\boldsymbol{x}}_n]^{\mathrm{T}}, \quad \boldsymbol{Y} = [\boldsymbol{y}_1 \cdots \boldsymbol{y}_q] = [\underline{\boldsymbol{y}}_1 \cdots \underline{\boldsymbol{y}}_n]^{\mathrm{T}},$$
$$\boldsymbol{W} = [\boldsymbol{w}_1 \cdots \boldsymbol{w}_q] = [\underline{\boldsymbol{w}}_1 \cdots \underline{\boldsymbol{w}}_m]^{\mathrm{T}}, \quad \boldsymbol{E} = [\boldsymbol{e}_1 \cdots \boldsymbol{e}_q] = [\underline{\boldsymbol{e}}_1 \cdots \underline{\boldsymbol{e}}_n]^{\mathrm{T}},$$

where each bolded lower-case letter refers to a column vector.

We estimate $\boldsymbol{W}^*$ by minimizing the error sum of squares subject to a sparsity constraint, namely by solving the L_2-SVS problem

$$\underset{\boldsymbol{W}}{\text{minimize}}\ f(\boldsymbol{W}) \quad \text{subject to}\ g(\boldsymbol{W}) \leqslant r, \quad \text{where}\ f(\boldsymbol{W}) = \tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_{\mathrm{F}}^2 \quad \text{and} \quad g(\boldsymbol{W}) = \sum_{j=1}^{m} \|\underline{\boldsymbol{w}}_j\|_2. \tag{2}$$