Computational Statistics and Data Analysis 124 (2018) 180–196

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Sparse principal component regression for generalized linear models

Shuichi Kawano [a,*], Hironori Fujisawa [b,c,d], Toyoyuki Takada [e], Toshihiko Shiroishi [e]

[a] Department of Computer and Network Engineering, Graduate School of Informatics and Engineering, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan
[b] The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
[c] Department of Statistical Science, The Graduate University for Advanced Studies, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
[d] Department of Mathematical Statistics, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, Nagoya, Aichi 466-8550, Japan
[e] Mammalian Genetics Laboratory, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan

ARTICLE INFO

Article history:
Received 28 August 2017
Received in revised form 8 March 2018
Accepted 10 March 2018
Available online 22 March 2018

Keywords:
Coordinate descent
Dimension reduction
Sparse regularization
Variable selection

ABSTRACT

Principal component regression (PCR) is a widely used two-stage procedure: principal component analysis (PCA), followed by regression in which the selected principal components are regarded as new explanatory variables in the model. Note that PCA is based only on the explanatory variables, so the principal components are not selected using the information on the response variable. We propose a one-stage procedure for PCR in the framework of generalized linear models. The basic loss function is based on a combination of the regression loss and PCA loss. An estimate of the regression parameter is obtained as the minimizer of the basic loss function with a sparse penalty. We call the proposed method sparse principal component regression for generalized linear models (SPCR-glm). Taking the two loss function into consideration simultaneously, SPCR-glm enables us to obtain sparse principal component loadings that are related to a response variable. However, a combination of loss functions may cause a parameter identification problem, but this potential problem is avoided by virtue of the sparse penalty. Thus, the sparse penalty plays two roles in this method. We apply SPCR-glm to two real datasets, doctor visits data and mouse consomic strain data.

© 2018 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Principal component regression (PCR) (Massy, 1965; Jolliffe, 1982) is a widely used two-stage procedure: one first performs principal component analysis (PCA) (Pearson, 1901; Jolliffe, 2002) and next considers a regression model in which the selected principal components are regarded as new explanatory variables. PCR has many extensions (Hartnett et al., 1998; Rosipal et al., 2001; Reiss and Ogden, 2007; Wang and Abbott, 2008). However, we should remark that PCA is based only on the explanatory variables, so the principal components are not selected using the information on the response variable. If the response variable has a close relationship with the principal components having small eigenvalues, PCR cannot achieve sufficient prediction accuracy.

* Corresponding author.
E-mail addresses: skawano@uec.ac.jp (S. Kawano), fujisawa@ism.ac.jp (H. Fujisawa), ttakada@nig.ac.jp (T. Takada), tshirois@nig.ac.jp (T. Shiroishi).

https://doi.org/10.1016/j.csda.2018.03.008
0167-9473/© 2018 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

To overcome this problem, Kawano et al. (2015) proposed a one-stage procedure for PCR. The basic loss function for this one-stage procedure is based on a combination of the regression squared loss and PCA loss (Zou et al., 2006). The estimate of the regression parameter is obtained as the minimizer of the basic loss function with a sparse penalty. This proposed method is called the sparse principal component regression (SPCR). SPCR enables us to obtain sparse principal component loadings that are related to a response variable, because the two loss functions are simultaneously taken into consideration. However, the response variable is restricted to a continuous variable.

In this paper, we propose a one-stage procedure for PCR in the framework of generalized linear models (McCullagh and Nelder, 1989). The regression loss is replaced by the negative log-likelihood function. The proposed method is called the sparse principal component regression for generalized linear models (SPCR-glm). The main difference in SPCR-glm from SPCR is the parameter estimation procedure, because the negative log-likelihood function in generalized linear models is more complex than the regression squared loss. To obtain the parameter estimate, we propose a novel update algorithm combining various ideas with the coordinate descent algorithm (Fu, 1998; Friedman et al., 2007; Wu and Lange, 2008).

The partial least squares (PLS) performs dimension reduction and regression analysis simultaneously (Wold, 1975; Frank and Friedman, 1993). This concept is similar to that in SPCR. In the framework of generalized linear models, Bastien et al. (2005) proposed PLS generalized linear regression (PLS-GLR). We will compare SPCR-glm with PLS-GLR numerically in Sections 5, 6, and 7.

SPCR-glm is applied to two real datasets, doctor visits data and mouse consomic strain data, with a Poisson regression model and multi-class logistic model, respectively. SPCR-glm provides more easily interpretable principal component (PC) loadings and clearer classification on PC plots than various competing methods. For the doctor visits data, we can obtain very clearly interpretable PC loadings. For the consomic strain mouse data, we can succeed to extract characteristic mouse consomic strains with smaller within-variance. Through simulation studies, we examine that SPCR-glm is superior or competitive to competing methods in terms of prediction accuracy.

This paper is organized as follows. In Section 2, we review sparse principal component analysis (SPCA) and SPCR. In Section 3, we propose SPCR-glm and introduce some special cases. In Section 4, we provide a parameter estimation procedure for SPCR-glm and discuss the selection of tuning parameters. Real data analyses and Monte Carlo simulations are illustrated in Sections 5, 6, and 7. Concluding remarks are given in Section 8. The R language software package **spcr**, which implements SPCR-glm, is available on the Comprehensive R Archive Network (R Core Team, 2018).

2. Preliminaries

2.1. Sparse principal component analysis

Let $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ be an $n \times p$ data matrix with n observations and p variables. Without loss of generality, the columns of the matrix X are assumed to be centered. PCA is formulated as the following least squares problem (e.g., Hastie et al. (2011)):

$$\min_{B} \sum_{i=1}^{n} \|\boldsymbol{x}_i - BB^T\boldsymbol{x}_i\|_2^2 \quad \text{subject to} \quad B^TB = I_k, \tag{1}$$

where $B = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ is a $p \times k$ principal component loading matrix, k denotes the number of principal components, I_k is the $k \times k$ identity matrix, and $\|\cdot\|_2$ is the L_2 norm defined by $\|\boldsymbol{z}\|_2 = \sqrt{\boldsymbol{z}^T\boldsymbol{z}}$ for an arbitrary finite vector $\boldsymbol{z}$. Let $X = UDV^T$, where U is an $n \times p$ matrix with $U^TU = I_p$, $V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$ is a $p \times p$ orthogonal matrix, and $D = \mathrm{diag}(d_1, \ldots, d_p)$ is a $p \times p$ matrix with $d_1 \geq \cdots \geq d_p \geq 0$. Then, the estimate of B is given by V_kQ^T, where $V_k = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ and Q is an arbitrary $k \times k$ orthogonal matrix.

To easily interpret the principal component loading matrix B, Zou et al. (2006) proposed SPCA, which is given by

$$\min_{A,B} \left\{ \sum_{i=1}^{n} \|\boldsymbol{x}_i - AB^T\boldsymbol{x}_i\|_2^2 + \lambda \sum_{j=1}^{k} \|\boldsymbol{\beta}_j\|_2^2 + \sum_{j=1}^{k} \lambda_{1,j}\|\boldsymbol{\beta}_j\|_1 \right\} \tag{2}$$
$$\text{subject to} \quad A^TA = I_k,$$

where $A = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k)$ is a $p \times k$ matrix, λ and the $\lambda_{1,j}$'s ($j = 1, \ldots, k$) are non-negative regularization parameters, and $\|\cdot\|_1$ is the L_1 norm defined by $\|\boldsymbol{z}\|_1 = \sum_{j=1}^{p}|z_j|$ for an arbitrary finite vector $\boldsymbol{z} = (z_1, \ldots, z_p)^T$. A simple calculation shows that SPCA can be expressed as

$$\min_{A,B} \sum_{j=1}^{k} \left\{ \|X\boldsymbol{\alpha}_j - X\boldsymbol{\beta}_j\|_2^2 + \lambda\|\boldsymbol{\beta}_j\|_2^2 + \lambda_{1,j}\|\boldsymbol{\beta}_j\|_1 \right\} \quad \text{subject to} \quad A^TA = I_k.$$

Given a fixed B, the minimizer A is obtained by solving the reduced rank Procrustes rotation, which is introduced in Zou et al. (2006). Given a fixed A, the minimization problem for B is consistent with that in the elastic net (Zou and Hastie, 2005),

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات