Neurocomputing 149 (2015) 1155–1161

Contents lists available at ScienceDirect

Neurocomputing

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

A probabilistic model for latent least squares regression

Shengzheng Wang [a,*], Jie Yang [b]

[a] Merchant Marine College, Shanghai Maritime University, Shanghai 201306, PR China
[b] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, PR China

ARTICLE INFO

Article history:
Received 15 April 2014
Received in revised form
8 August 2014
Accepted 6 September 2014
Communicated by D. Tao
Available online 16 September 2014

Keywords:
Least squares regression
Structural information
Latent least squares regression
Pattern recognition

ABSTRACT

By far, least squares regression (LSR) is the most widely used data modeling method in statistics and mathematics because of its effectiveness and completeness. It plays an important underlying role in many extensions, e.g., regularized LSR, weighted LSR, and *lasso*. Since LSR is a discriminative model, it allows only sampling of the target variables conditioned on observations. In this paper, we present the latent LSR (LLSR), a generative model, which enables LSR to exploit the structural information hidden in the explanatory variables by imposing a sparsity-encouraging prior over the precision matrix of the latent variable. A maximum a posteriori (MAP) estimate is applied to obtain a point estimate of the model parameters. Both the toy example and real data tests suggest the effectiveness of LLSR.

1. Introduction

Least squares regression (LSR), the most popular linear model in statistics and mathematics, is an effective tool in pattern recognition and machine learning. Given a set of training samples, LSR finds the coefficients of a linear model by minimizing the residual sum of squares. When the empirically estimated covariance matrix is of full rank, the model coefficients of least squares can be efficiently obtained by solving a linear system. LSR achieves competitive performance in many practical applications compared with sophisticated nonlinear models, especially when the amount of training samples is small or the signal-to-noise ratio is low [1].

LSR plays an important underlying role in many extensions, which can be frequently founded in the literature, e.g., L2 norm regularized LSR (RLSR) [2], locally regularized LSR [3], entropy regularized LSR [4], weighted LSR [5], partial least square regression (PLSR) [6], orthogonal LSR [7], *lasso* [8], and some recent works including [9–11], and [12]. Among them, RLSR and *lasso* are very popular. In particular, RLSR can alleviate the over-fitting phenomena in the least squares by imposing an ℓ_2 norm regularization on the model coefficients. According to the statistical learning theory, this regularization improves the generalization by restricting the volume of the solution space. By replacing the ℓ_2 norm in the regularization with the ℓ_1 norm, RLSR becomes *lasso*. The ℓ_1 norm regularization is equivalent to imposing a Laplacian prior on the regression coefficients, and thus encourages the sparsity of the model coefficients. In *lasso*, coefficients slightly correlated to responses shrink to zero, while coefficients strongly correlated to responses are retained. This sparsity makes the learned model more succinct and simpler; controls the weights of original variables and decreases the variance brought by possible over-fitting with the least increment of the bias; and provides a good interpretation of the model to reveal an explicit relationship between the objective of the model and the given variables.

From the viewpoint of machine learning, however, LSR and its popular extensions are discriminative models, which are equivalent to maximizing their corresponding conditional likelihood functions. Thus, LSR and its extensions cannot exploit the hidden structural information in the training samples. In this paper, we assume that the explanatory variables are contaminated with noises, and the latent distribution of the explanatory variables has structural information encoded by a sparse precision matrix. In particular, the resulting latent least square regression (LLSR) models both the generation of the explanatory variables and the responses via the latent variable. We carefully design an expectation–maximization (EM) algorithm for finding the maximum a posteriori (MAP) estimates of model parameters. The proposed LLSR enjoys the following advantages:

- It is a generative model, and thus is less sensitive to over-fitting problem even when the amount of training samples is relatively small.
- It exploits the structural information underlying the precision matrix of the latent variables by using a sparsity-encouraging prior. This significantly reduces the inaccurate estimation of the covariance in LSR.
- The tuning parameters are integrated out by using the Bayesian inference, and thus no cross validation is required.

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2014.09.014
0925-2312/

The rest of this paper is organized as follows. In Section 2, after a brief introduction of LSR, we present the proposed LLSR. In Section 3, we derive an EM type algorithm to learn the model parameters. Section 4 presents a toy example to demonstrate the effectiveness of LLSR in improving the estimation of precision matrix and the generalization ability. Section 5 compares LLSR against LSR and RLSR on three benchmark datasets. Section 6 concludes this paper (Table 1).

Table 1
Summary of important notations throughout this paper.

Notation	Description
$\{x, y, z\}$	Explanatory variable, response, and latent variable
$\{\mu, S\}$	Parameters for z, i.e., mean and precision matrix
σ_1	Noise level, i.e., standard deviation, for x
σ_2	Noise level, i.e., standard deviation, for y
$\{w, b\}$	Regression parameter
Θ	All parameters $\{\mu, S, w, b, \sigma_1, \sigma_2\}$
$p(x, y\|\Theta)$	Joint likelihood
$p(\Theta)$	Prior probability, for Θ
$\hat{a}$	An empirical estimation of variable a
$KL(q\|\|p)$	Kullback–Leibler divergence between distribution q and p

2. Latent least squares regression

In this section, we first briefly review LSR. Since it is equivalent to the maximization of conditional likelihood, LSR is unable to exploit structural knowledge hidden in data. Then, we present LLSR, which is motivated by addressing this limitation of LSR.

2.1. Least squares regression

A linear regression function $y = f(x)$ models the relationship between the explanatory variables denoted $x \in \mathbb{R}^m$ and the response variable denoted $y \in \mathbb{R}$, such that the model depends linearly on the unknown parameters to be estimated from the data. The regression function $f(x)$ takes the form

$$f(x) = b + w^T x, \tag{1}$$

where $b \in \mathbb{R}$ is the bias and $w \in \mathbb{R}^m$ is the parameter vector that contains the regression coefficients [1]. Given a set of training samples $\{(x_1, y_1), (x_2, _2), \ldots, (x_n, y_n)\}$, LSR estimates b and w by minimizing the residual sum of squares

$$RSS = \sum_{i=1}^{n} (y_i - b - w^T x_i)^2. \tag{2}$$

By assuming a conditional Gaussian distribution on the response y, i.e., $y \sim \mathcal{N}(f(x), \sigma^2)$, the minimization defined by Eq. (2) has a probabilistic interpretation that the log conditional likelihood on n independent and identically distributed training samples takes the form

$$\mathcal{L} = -n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(x_i))^2 + Const. \tag{3}$$

It is direct to obtain that the maximization of Eq. (3) is equivalent to the minimization of Eq. (2), although the former provides the estimate of σ. Fig. 1 shows LSR cannot model explanatory variables x. Therefore, LSR has the following shortcomings. First, it is unable to deal with the case where the explanatory variables x are contaminated with noises. Second, it does not utilize any structural information hidden in the explanatory variables x. And in this study, we consider one certain type of structural information, i.e., the sparsity of the precision matrix for the latent distribution of explanatory variables.

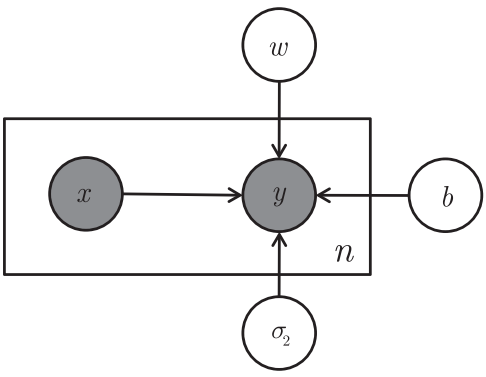

Fig. 1. The graphical model representation of LSR.

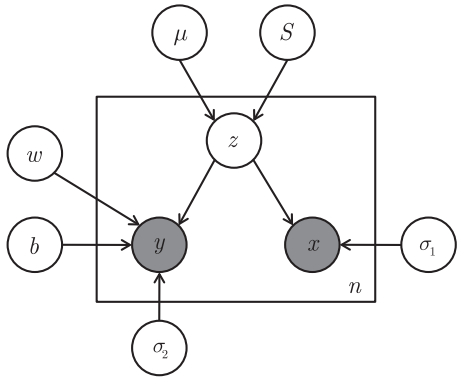

Fig. 2. The graphical model representation of LLSR.

2.2. Latent least squares regression

In latent least squares regression (LLSR), a latent variable z is introduced, and the explanatory variables x and the response y are modeled as

$$\begin{cases} x = z + e_1 \\ y = b + w^T z + e_2, \end{cases} \tag{4}$$

where we assume that $z \in \mathbb{R}^m$ is sampled from a Gaussian distribution $\mathcal{N}(\mu, S^{-1})$,[1] while noises $e_1 \in \mathbb{R}^m$ and $e_2 \in \mathbb{R}$ are sampled from $\mathcal{N}(0, \sigma_1 I)$ and $\mathcal{N}(0, \sigma_2)$, respectively. The probability distributions of z, x, and y are given by

$$p(z|\mu, S) = \frac{\det(S)^{1/2}}{(2\pi)^{m/2}} e^{-(1/2)(z-\mu)^T S(z-\mu)}, \tag{5}$$

$$p(x|z, \sigma_1) = \frac{1}{(\sqrt{2\pi}\sigma_1)^m} e^{-(1/2\sigma_1^2)(x-z)^T(x-z)}, \tag{6}$$

$$p(y|z, w, b, \sigma_2) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-(1/2\sigma_2^2)(y-w^T z-b)^2}. \tag{7}$$

Fig. 2 is the graphical model representation of LLSR and shows that LLSR models both the generation of the explanatory variables x and the response y by introducing a latent variable z. In the next subsection, we show that the precision matrix S facilitates the exploration of structural information hidden in the explanatory variables.

Let the model parameters of LLSR be $\Theta = \{\mu, S, w, b, \sigma_1, \sigma_2\}$, and then the complete joint probability of (z, x, y) is given by

$$p(z, x, y|\Theta) = p(z|\mu, S)p(x|z, \sigma_1)p(y|z, w, b, \sigma_2). \tag{8}$$

[1] The S is the precision matrix, which is the inverse of the covariance matrix.