Computational Statistics and Data Analysis 68 (2013) 190–201

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Kernel continuum regression☆

Myung Hee Lee [a,*], Yufeng Liu [b]

[a] Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA

[b] Department of Statistics and Operations Research, Carolina Center for Genome Sciences, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA

ARTICLE INFO

Article history:
Received 3 August 2012
Received in revised form 16 June 2013
Accepted 17 June 2013
Available online 1 July 2013

Keywords:
Continuum regression
Kernel regression
Ordinary least squares
Principal component regression
Partial least squares

ABSTRACT

The continuum regression technique provides an appealing regression framework connecting ordinary least squares, partial least squares and principal component regression in one family. It offers some insight on the underlying regression model for a given application. Moreover, it helps to provide deep understanding of various regression techniques. Despite the useful framework, however, the current development on continuum regression is only for linear regression. In many applications, nonlinear regression is necessary. The extension of continuum regression from linear models to nonlinear models using kernel learning is considered. The proposed kernel continuum regression technique is quite general and can handle very flexible regression model estimation. An efficient algorithm is developed for fast implementation. Numerical examples have demonstrated the usefulness of the proposed technique.

1. Introduction

Regression is one of the most fundamental and useful statistical techniques. It helps to relate explanatory variables with a response variable and build predictive models. Ordinary Least Squares (OLS) regression estimates the conditional mean of the response variable given covariates and is commonly used in practice. Despite its simple implementation and good interpretability, OLS may face numerical unstability when there exists multicollinearity among covariates or when the dimension of covariates is relatively high. In that case, Ridge Regression (RR), which can be viewed as a penalized approach, may serve as an alternative.

Another popular group of regression techniques is to perform regression analysis based on a small number of linear transformations of the explanatory variables. For example, Principal Component Regression (PCR) first summarizes multiple explanatory variables, which can be high dimensional, into a few principal component directions and then performs regression on those principal component directions. These principal component directions are orthogonal to each other, yet contain most of the variations in the explanatory variables. Thus, PCR can circumvent the potential numerical difficulty of OLS. Partial Least Squares (PLS) is a related regression technique and it has been widely used in the field of chemometrics. Similar to PCR, PLS also uses a small number of linear transformations of the covariates for regression. The main difference of PLS from PCR is that PCR finds those transformations without the use of the response variable while PLS makes use of both covariates and the response variable to seek for suitable transformations.

With various regression techniques available, it would be desirable to study the differences and connections among these methods. Stone and Brooks (1990) pointed out that the seemingly different regression procedures such as OLS, PLS,

☆ This article has supplementary material online (see the Appendix).

* Corresponding author. Tel.: +1 970 491 6682; fax: +1 970 491 7895.
E-mail addresses: mhlee@stat.colostate.edu (M.H. Lee), yfliu@email.unc.edu (Y. Liu).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.06.016

and PCR differ only in one aspect: the target quantity maximized at the first step when finding linear transformations of the explanatory variables. Based on this analogy, they formulated a richer family of regression methods, Continuum Regression (CR), by introducing a continuum parameter which connects these three methods.

Similar to other methods, CR also aims to find directional vectors to transform the explanatory variables into new latent predictors which are orthogonal to each other and are constructed as linear combinations of the original predictors. There are two aspects of the latent predictors: one is the variation of the original predictors explained by each latent predictor and the other is the correlation between each latent predictor with the response. The quantity for the CR to maximize involves both variance and correlation of the latent variable with a parameter controlling the relative proportion of two terms. With its flexible construction, the CR is quite general and it contains OLS and PCR as the two extremes and PLS in the middle. In particular, the OLS ignores the variance of the latent variable and maximizes the correlation between the observed response vector and the predicted response vector. In contrast, PCR finds the regression directional vector so that the variance of the latent predictor is maximized. Interestingly, PLS essentially maximizes the covariance between the observed and the predicted response vectors. Besides these three special cases, CR also covers many other methods in the whole spectrum. Frank and Friedman (1993) provides a nice overview of CR and other related regression techniques. Sundberg (1993) and Björkström and Sundberg (1999) reveal some close connection between the RR and the set of CR. Chen and Cook (2010) investigated some asymptotic properties of CR.

The CR approach is potentially useful when the relationship between the response and the explanatory variable is linear. However, when the true relationship is nonlinear, the predictive performance of the CR family can be improved if the model is built as a nonlinear function of the explanatory variables.

In the literature, there has been some work in this direction which generalizes some special cases of CR via kernel learning. The nonlinear generalization of the building blocks for PCR, i.e., nonlinear Principal Component Analysis (PCA) has been studied in the field of pattern recognition, where lower dimensional feature extraction of high dimensional data becomes an important task. See Schölkopf et al. (1998), Mika et al. (1999), and Shawe-Taylor and Cristianini (2004) for more details. Rosipal et al. (2000b) and Rosipal et al. (2000a) deal with nonlinear generalization of PCR, which uses nonlinear PCA as the latent variables in the regression analysis. As in the linear case, the feature extraction based on PCA is done not specifically to the regression problem at hand, and consequently the predictive performance of PCR is usually not as good as PLS. Walczak and Massart (1996) and Rosipal and Trejo (2001) generalize the PLS to incorporate nonlinear cases.

In this present paper, we extend the linear CR model to nonlinear CR model using the powerful kernel trick concept in machine learning. The proposed kernel CR (kCR) incorporates the special cases such as kernel OLS, kernel PLS and kernel PCR in one unified framework. If a linear kernel map is chosen, the kCR is the same as the ordinary CR. Section 2 provides mathematical formulation of kCR. The first part is devoted to present systematic ways to construct latent variables from an optimization point of view, and the second part is to run a regression analysis with the selected latent variables. Section 3 gives the details of the algorithm for solving the optimization problem in the first step. Numerical performance of the proposed method is investigated in Section 4 through simulation examples and real data analysis. We conclude the paper with some brief discussion in Section 5.

2. Continuum regression and its kernel extension

In this section, we first briefly review the linear CR in Section 2.1 and then introduce its kernel extension in Section 2.2.

2.1. Review of linear continuum regression

Suppose that we have n pairs of data points for regression, $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ are the explanatory variables and $y_i \in \mathbb{R}$ is the response variable. Define the $n \times d$ input data matrix as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}}$, where each row vector $\mathbf{x}_i$ represents a d-dimensional input vector for $i = 1, \ldots, n$. The output data vector is denoted by $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$. We assume that the data are mean-centered so that each column sum of the matrix $\mathbf{X}$ is $\mathbf{0}$. Denote X and Y as the random predictor vector and response variable respectively. Furthermore, define the scatter matrix of the data $\mathbf{X}$ as $\mathbf{S}_{d\times d} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ and the cross covariance matrix between $\mathbf{X}$ and $\mathbf{y}$ as $\mathbf{s} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$.

We now describe the linear CR technique in terms of its optimization criterion. CR is essentially a two-step regression procedure where in the first step, one finds a set of direction vectors in the input variable space, $\mathbb{R}^d$, and makes projections of data onto the subspace generated by these vectors. In the second step, we use these extracted features as regressors to build a regression model to predict the value of the response variable Y. In particular, for a given parameter $\alpha \in [0, 1]$, suppose that the first k direction vectors $\mathbf{c}_1, \ldots, \mathbf{c}_k$ have been constructed and we want to find $\mathbf{c}_{k+1}$ by maximizing

$$\begin{aligned} T(\mathbf{c}) &= \mathrm{Cov}(\mathbf{c}^{\mathrm{T}}X, Y)^2 \, \mathrm{Var}(\mathbf{c}^{\mathrm{T}}X)^{\alpha/(1-\alpha)-1} \\ &= (\mathbf{c}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y})^2 (\mathbf{c}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{c})^{\alpha/(1-\alpha)-1} \\ &= (\mathbf{c}^{\mathrm{T}}\mathbf{s})^2 (\mathbf{c}^{\mathrm{T}}\mathbf{S}\mathbf{c})^{\alpha/(1-\alpha)-1} \end{aligned} \tag{1}$$

subject to the constraints $\|\mathbf{c}\|^2 = 1$ and $\mathrm{Corr}(\mathbf{c}^{\mathrm{T}}X, \mathbf{c}_j^{\mathrm{T}}X) = \mathbf{c}^{\mathrm{T}}\mathbf{S}\mathbf{c}_j = 0, j = 1, \ldots, k$.

Stone and Brooks (1990) showed that CR includes OLS and PCR at the two extremes, $\alpha = 0$ and $\alpha \to 1$, respectively. Specifically, OLS can be viewed as maximizing correlation. In particular, the multiple correlation coefficient is maximized