ELSEVIER

Biochemical Engineering Journal 40 (2008) 199–204

Biochemical Engineering Journal

www.elsevier.com/locate/bej

Canonical correlation analysis for multivariate regression and its application to metabolic fingerprinting

Hiroyuki Yamamoto [a], Hideki Yamaji [b], Eiichiro Fukusaki [c], Hiromu Ohno [b], Hideki Fukuda [a,d,*]

[a] *Department of Molecular Science and Material Engineering, Graduate School of Science and Technology, Kobe University, 1-1 Rokkodai, Nada, Kobe 657-8501, Japan*

[b] *Department of Chemical Science and Engineering, Graduate School of Engineering, Kobe University, 1-1 Rokkodai, Nada, Kobe 657-8501, Japan*

[c] *Department of Biotechnology, Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita 565-0871, Japan*

[d] *Organization of Advanced Science and Technology, Kobe University, 1-1 Rokkodai, Nada, Kobe 657-8501, Japan*

Received 27 July 2007; received in revised form 10 December 2007; accepted 15 December 2007

Abstract

Multivariate regression analysis is one of the most important tools in metabolomics studies. For regression of high-dimensional data, partial least squares (PLS) has been widely used. Canonical correlation analysis (CCA) is a classic method of multivariate analysis; it has however rarely been applied to multivariate regression. In the present study, we applied PLS and regularized CCA (RCCA) to high-dimensional data where the number of variables (p) exceeds the number of observations (N), $N \ll p$. Using kernel CCA with linear kernel can drastically reduce the calculation time of RCCA. We applied these methods to gas chromatography–mass spectrometry (GC–MS) data, which were analyzed to resolve the problem of Japanese green tea ranking. To construct a quality-predictive model, the optimal number of latent variables in RCCA determined by leave-one-out cross-validation (LOOCV) was significantly fewer than in PLS. For metabolic fingerprinting, we successfully identified important metabolites for green tea grade classification using PLS and RCCA.

Keywords: Canonical correlation analysis; Partial least squares; Kernel method; Multivariate analysis; Metabolic fingerprinting; Metabolomics

1. Introduction

Metabolomics is a science based on exhaustive-profiling of metabolites. It has been widely applied to animals and plants, microorganisms, food and herbal medicine materials, and other areas. In metabolomics, gas chromatography–mass spectrometry (GC–MS), liquid chromatography–mass spectrometry (LC–MS), and capillary electrophoresis–mass spectrometry (CE–MS) are all important technologies for the analysis of metabolites [1]. Metabolic fingerprinting [2,3] is a technology that considers the metabolome to be a fingerprint and is applied to various classifications and forecasts. The procedures include the identification of important metabolites for regression or classification by applying multivariate analysis or machine learning to data obtained by the above-mentioned analytical methods.

Several multivariate regression methods have been applied in metabolomics studies [4,5]. For regression and classification of high-dimensional data, partial least squares (PLS) [6,7] have been widely used so far. Recently, PLS has been used in the field of bioinformatics research to analyze gene expression data from cDNA microarrays [8,9]. The main reason why PLS has been widely used is its ready applicability where the number of variables (p) exceeds the number of observations (N), $N \ll p$, and where there is multicolinearity among the variables.

Canonical correlation analysis (CCA) [10] is, like principal component analysis (PCA), a classic method of multivariate analysis; it is however rarely applied to high-dimensional data for regression because it is theoretically impossible to apply CCA to $N \ll p$ type data, to which we can however apply regularized CCA (RCCA). The value of the regularized parameter in RCCA interpolates smoothly between PLS and CCA [11].

The kernel method [12,13] has been studied mainly in machine learning since a support vector machine was developed and actively studied in the field of bioinformatics research [14]. Nonlinear extension of multivariate analysis using the kernel

* Corresponding author at: Organization of Advanced Science and Technology, Kobe University, 1-1 Rokkodai, Nada, Kobe 657-8501, Japan.
Tel.: +81 78 803 6192; fax: +81 78 803 6192.
E-mail address: fukuda@kobe-u.ac.jp (H. Fukuda).

1369-703X/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.bej.2007.12.009

method, including kernel PCA [15], kernel Fisher discriminant analysis (FDA) [16], kernel PLS [17], and kernel CCA [18,19], has been proposed. We can perform nonlinear multivariate analysis by replacing the inner products in the feature space with the kernel function without explicitly knowing the mapping in the feature space.

In the present study, we applied PLS and RCCA to GC–MS data, which were analyzed to resolve the problem of Japanese green tea ranking. The main objective of the present study is to apply RCCA to $N \ll p$ type data and compare RCCA with PLS. When we apply an ordinary PLS algorithm to large-size data such as high-dimensional data, the algorithm often requires a large amount of memory and long computational time. An alternative PLS algorithm to avoid these problems is therefore proposed [20]. These problems are more serious in RCCA than in PLS because of the need to handle a large-size matrix, $p \times p$. Using kernel CCA with a linear kernel allows the use of a small size matrix, $N \times N$.

2. Data analysis

2.1. Data

In the present study, we used data from GC–MS in which hydrophilic primary green tea metabolites were analyzed [21]. The main purpose of the Japanese green tea ranking problem is to construct a quality-predictive model. Data preprocessing including peak alignment, peak identification, and conversion to numeric variables was achieved in a way similar to that previously reported [21]. The explanatory variable $\mathbf{X}$ consists of metabolite-profiling data from chromatography. The response variable $\mathbf{y}$ is ranking of teas from 1st to 53rd determined by the total scores of the sensory tests, which are leaf appearance, smell, and color of the brew and its taste, judged by professional tea testers. The explanatory variable $\mathbf{X}$ and the response variable $\mathbf{y}$ are mean-centered but are not scaled. Fifty-three samples were divided into 2 groups: 47 samples as a training set and 6 samples, those ranked 2nd, 12th, 22nd, 32nd, 42nd, and 52nd, excluded as a test set. Each data set contained 2064 variables in which retention time changed every 0.01 min from 4.01 to 24.64 min.

2.2. Data analysis methods

Multiple linear regression (MLR) is an ordinary regression analysis; it constructs a regression model between the explanatory variable $\mathbf{X}$ and the response variable $\mathbf{y}$. However, MLR cannot be applied to $N \ll p$ type data. Regression methods by using latent variables such as PLS construct a regression model between a new explanatory variable $\mathbf{t}$, which is obtained by dimensionality reduction of $\mathbf{X}$, and the response variable $\mathbf{y}$. Here we explain the dimensionality reduction method in PLS, CCA, RCCA, kernel PLS, and kernel CCA as a generalized eigenvalue problem, as described previously [22].

2.2.1. Partial least squares

PLS is explained as the optimization problem of maximizing the square of covariance between the score vector $\mathbf{t}$, which is a linear combination of the explanatory variable $\mathbf{X}$, and the response variable $\mathbf{y}$ under the constraint of $\mathbf{w}'\mathbf{w} = 1$:

$$\begin{aligned} &\max \ [\mathrm{cov}(\mathbf{Xw}, \mathbf{y})]^2 \\ &\text{const.} \ \ \mathbf{w}'\mathbf{w} = 1 \end{aligned}$$

where $\mathbf{w}$ is a weight vector. $\mathbf{X}$ and $\mathbf{y}$ are mean-centered. Finally, PLS is formulated as the following eigenvalue problem:

$$\frac{1}{N^2}\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}\mathbf{w} = \lambda\mathbf{w} \tag{1}$$

where λ is a Lagrange multiplier.

The eigenvector corresponding to the maximum eigenvalue is the weight vector of PLS. This eigenvalue problem is solved by singular value decomposition (SVD). A score vector can be calculated as $\mathbf{t} = \mathbf{Xw}$. To calculate more than one latent variable, we perform deflation of $\mathbf{X}$ and $\mathbf{y}$ and then calculate the eigenvector corresponding to the maximum eigenvalue in Eq. (1). This operation is iterated until the number of latent variables reaches the required number.

2.2.2. Canonical correlation analysis

CCA is explained as the optimization problem of maximizing the square of correlation between the score vector $\mathbf{t}$, which is a linear combination of the explanatory variable $\mathbf{X}$, and the response variable $\mathbf{y}$:

$$\max \ [\mathrm{corr}(\mathbf{Xw}, \mathbf{y})]^2 = \left[\frac{\mathrm{cov}(\mathbf{Xw}, \mathbf{y})}{\sqrt{\mathbf{w}'\mathbf{X}'\mathbf{Xw}/N}}\right]^2$$

This conditional equation is rewritten as follows:

$$\begin{aligned} &\max \ [\mathrm{cov}\,(\mathbf{Xw}, \mathbf{y})]^2 \\ &\text{const.} \ \ \frac{1}{N}\mathbf{w}'\mathbf{X}'\mathbf{Xw} = 1 \end{aligned}$$

Finally, CCA is formulated as the following generalized eigenvalue problem:

$$\frac{1}{N}\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}\mathbf{w} = \lambda\mathbf{X}'\mathbf{X}\mathbf{w} \tag{2}$$

This generalized eigenvalue problem is solved by Cholesky decomposition of $\mathbf{X}'\mathbf{X}$ when $\mathbf{X}'\mathbf{X}$ is full rank and SVD.

2.2.3. Regularized canonical correlation analysis

In contrast to PLS, CCA is not applicable to the case $N \ll p$ because the matrix $\mathbf{X}'\mathbf{X}$ is rank-deficient. A penalty on the norm of the weight vector is introduced into CCA. This RCCA is applicable to the case $N \ll p$ because $\mathbf{X}'\mathbf{X} + \tau\mathbf{I}$ is always a full rank matrix. $\mathbf{I}$ denotes the identity matrix and τ the regularized parameter.

$$\begin{aligned} &\max \ \ [\mathrm{cov}(\mathbf{Xw}, \mathbf{y})]^2 \\ &\text{const.} \ \ (1-\tau)\frac{1}{N}\mathbf{w}'\mathbf{X}'\mathbf{Xw} + \tau\mathbf{w}'\mathbf{w} = 1 \end{aligned}$$