MULTIVARIATE ANALYSIS APPROACH FOR METABOLOME DATA ANALYSIS

Yamamoto, Hiroyuki

(Degree)
博士（工学）

(Date of Degree)
2008-03-25

(Date of Publication)
2012-02-20

(Resource Type)
doctoral thesis

(Report Number)
甲4199

(URL)
https://hdl.handle.net/20.500.14094/D1004199

神戸大学 博士論文

MULTIVARIATE ANALYSIS APPROACH FOR METABOLOME DATA ANALYSIS

多変量解析を用いたメタボロームデータ解析に関する研究

平成 20 年 1 月

神戸大学大学院自然科学研究科

山本　博之

ACKNOWLEDGEMENT

This is a thesis submitted by the author to Kobe University for the degree of Doctor of Engineering. The studies collected here were carried out between 2005 and 2008 under the direction of Professor Hideki Fukuda at the Biochemical Engineering Laboratory, Division of Molecular Science and Material Engineering, Graduate School of Science and Technology, Kobe University.

First of all, the author expresses the sincerest gratitude to research advisor, Professor Hideki Fukuda, for his continuous guidance and invaluable suggestions, and encouragement in the course of the studies. Next, the author expresses hearty gratitude to Professor Akihiko Kondo for his enormously beneficial discussion and kind support. The author deeply grateful to Professor Yasukiyo Ueda, Professor Eiichiro Fukusaki (Osaka University), Associate Professor Hideki Yamaji, Assistant Professor Tomohisa Katsuda, Assistant Professor Tsutomu Tanaka for their informative advice and hearty encouragement through the work.

The author send his utmost gratitude to Dr. Hiromu Ohno for giving an opportunity to further studies in his desired area and invaluable suggestions. The author further pays his acknowledgement to Dr. Satoshi Katahira (Toyota Central R&D), Dr. Shinji Hama (Bioenergy Corp.), Dr. Sriappareddy Tamalampudi, Mr. Keishi Hada (Otsuka Chemical Corp.), Mr. Naoki Shindo and all the members of Professor Fukuda's laboratory for their technical assistance and encouragement.

Last but not least, the author expresses deep appreciation to his parents, Masayuki and Taeko Yamamoto for their constant assistance and financial support, and grandfather Kametaro Ikeda for his encouragemnt.

Hiroyuki Yamamoto

Biochemical Engineering Laboratory

Division of Molecular Science and Material Engineering

Graduate School of Science and Technology

Kobe University

CONTENTS

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]. Especially, CE-MS technique has some advantageous points for metabolomics. It is suited to detect charged species, so comprehensive and simultaneous analysis of several anionic metabolites can be achieved [2, 3].

A statistical approach such as **multivariate analysis**, and information science technique is essential for metabolomics research. The combination of metabolomics, biochemical methodology, and informatics technique (bioinformatics) has the great potentiality in post genomic era. Soga et al [4] found that ophthalmic acid is a new biomarker of acetaminophen-induced hepatotoxicity by using metabolome analysis with differential analysis developed by Baran et al [5]. Functional genomics, which reveals biochemical functions corresponding to DNA sequence, by using metabolomics has been reported in many publications [6, 7, 8, 9]. Hirai et al [10, 11] identified a group of genes concerned with glucosinolate biosynthesis by analyzing metabolome and transcriptome with batch learning self-organizing map (BL-SOM) developed by Kanaya et al [12]. Prediction of the retention time of unknown metabolites in chromatography with computational method is also challenging problem [13, 14].

Data from analytical instruments, such as GC-MS, LC-MS, and CE-MS, are usually high-dimensional data. **Dimensionality reduction** technique is required as a preprocessing step for multivariate regression, visualization, and discrimination. Curve resolution problem, which resolve the overlapping peaks in chromatogram, is useful for metabolomics because some metabolites are often coeluted in chromatogram. This problem is also interpreted as dimensionality reduction under non-negativity

constraints.

Dimensionality reduction for multivariate regression

The regression problem [15] is to construct regression model between the explanatory variable and response variable. A multiple linear regression (MLR) is ordinary regression method, however, it is difficult to apply for high-dimensional data by multicollinearity. Especially, MLR cannot be applied for the data in which the number of variables (p) exceeds the number of observation (N), $N << p$. A variables selection by Akaike information criteria (AIC) [16] or t-statistic sometimes used to reduce the number of variables (model selection). It is not practical approach for high-dimensional and $N << p$ type data.

In part I, dimensionality reduction technique for multivariate regression is described. Fig. 1 shows the scheme of the dimensionality reduction for multivariate regression.

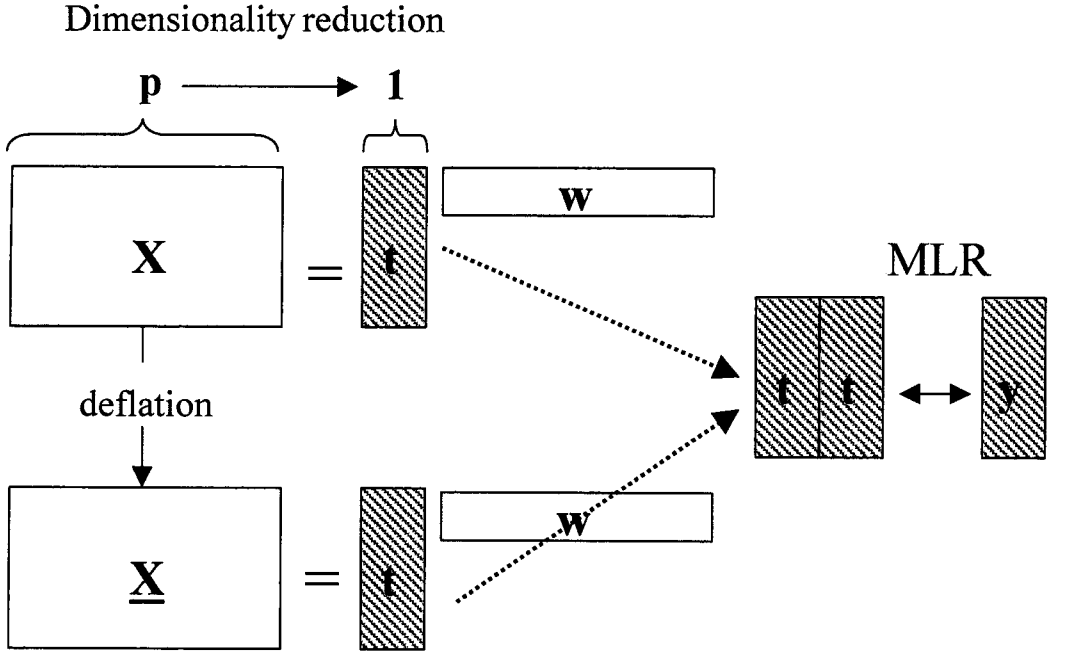

Fig. 1. Multivariate analysis for regression

X is the original data, and **X** is deflated to $\underline{\mathbf{X}}$. **t** is a latent variable, and **w** is the weight vector. The dimension of the original data is reduced from p to 1 by using dimensionality reduction, and achieved deflation of **X**. This operation has been iterated since the required number of latent variables is computed.

Dimensionality reduction for visualization and discrimination

Visualization by projection onto 2 or 3 dimensional subspace helps us to understand the data structure. And discrimination in high-dimensional space often causes the bad accuracy of prediction.

In part II, dimensionality reduction technique for visualization and discrimination is described.

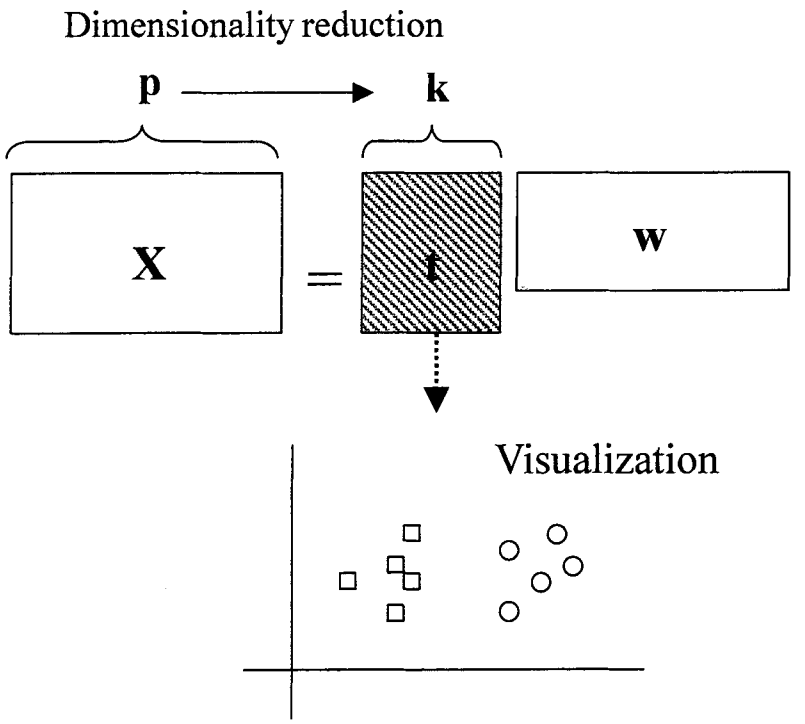

Fig. 2. Multivariate analysis for visualization and discrimination

Fig. 2 shows the scheme of the dimensionality reduction for visualization and discrimination. The dimension of the original data is reduced from p to k by using dimensionality reduction. Unsupervised dimensionality reduction by principal component analysis (PCA) can extract n latent variables in the case $N << p$, otherwise, supervised dimensionality reduction by partial least squares (PLS) and Fisher discriminant analysis (FDA) can extract only k = c (the number of class)-1 latent variables.

Dimensionality reduction for curve resolution

Curve resolution problem has been mainly studied in chemometrics [17] to resolve the overlapping peaks in data from analytical instrument, such as chromatography and near-infrared spectroscopy. It has common problem with separation of mixed signal in signal processing and feature extraction of face recognition in image processing.

In part III, dimensionality reduction technique for curve resolution is described. Fig. 3 shows the scheme of the dimensionality reduction for curve resolution. The

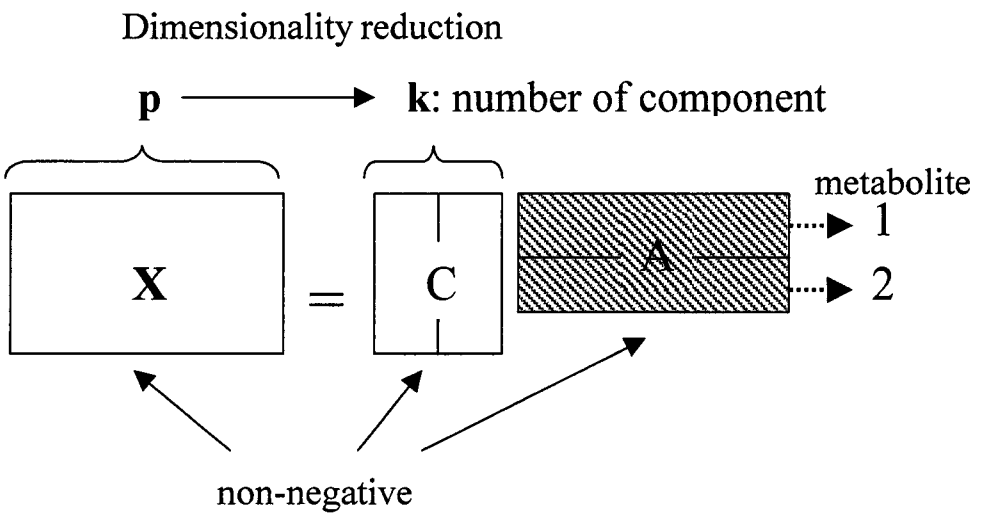

Fig. 3. Multivariate analysis for curve resolution

dimension of the original data is reduced from p to k (number of pure metabolites) by using dimensionality reduction under non-negative constraints. The original data is decomposed to the concentration matrix C and A which includes each pure components (metabolites) in each rows.

An overview of multivariate analysis

A multivariate analysis such as PCA, PLS, Fisher discriminant analysis (FDA) [18], have been applied for dimensionality reduction. PCA has been widely used in many research areas. PLS was developed by Wold [19] and mainly studied in chemometrics research area. FDA has been applied for pattern recognition, such as handwriting recognition.

The results of toy examples are shown in Fig. 4 Sample data (A), the results of PCA (B), PLS (C), and FDA (D) are shown respectively.

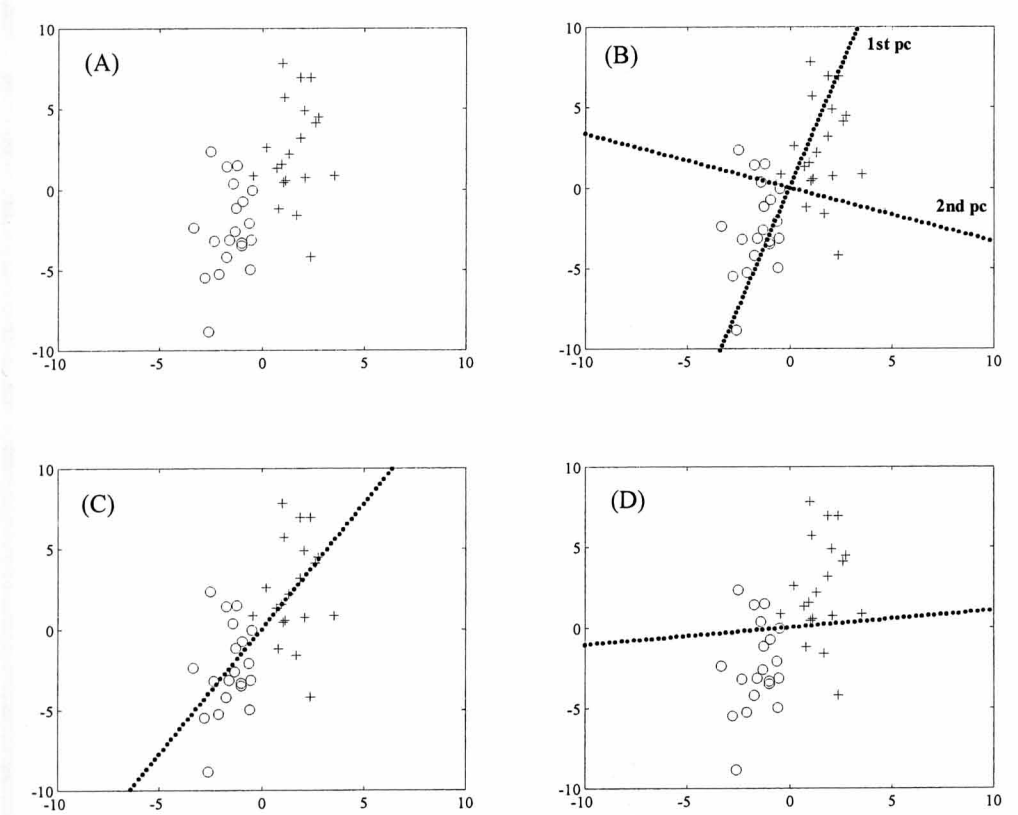

Fig. 4. Results of toy problem

In Fig.4, the symbols (○, +) belong to the each class. The dotted line denotes the 1-dimensional subspace obtained by each method. We want to find the direction of projection that achieved high separation of each class (binary classification problem). A direction of 1st axis in PCA (Fig.4 (B)) denotes the direction of the maximum variance of whole data, and 2nd axis is orthogonal to 1st axis. PLS and FDA find the direction which considers the separation between class. The axis of PLS (Fig.4 (C)) denotes the direction between the centers of each class. The axis of FDA (Fig.4 (D)) denotes the direction that maximizes between the centers of each class and minimizes the scatter of observation within each class.

Several dimensionality reduction methods have been proposed and ordinary multivariate analysis was extended. For example, locality-preserving projections (LPP) [20] optimally preserves the neighborhood structure of the data by using graph Laplacian. Local Fisher discriminant analysis (LFDA) [21] was developed for the supervised version of LPP. LFDA shows good results for the data in which the distribution of the observations within class is multimodal. An extension of PLS and FDA are also reported [22, 23]. Recently, nonlinear extension of linear classifier and multivariate analysis by kernel method are proposed. Vapnik [24] first introduced kernel methods to linear classifier as support vector machine. The main advantageous point of nonlinear extension by kernel methods is easy to compute by "kernel trick". Nonlinear multivariate analysis by kernel methods, such as kernel PCA [25], kernel PLS [26], kernel FDA [27], were proposed, and kernel methods are applied in bioinformatics [28]. Books about detailed theory of kernel methods have been published [29, 30].

An overview of curve resolution problem

Curve resolution problem, which resolve the overlapping peaks in chromatogram and near-infrared spectroscopy, has been studied in chemometrics research area. This problem can be interpreted to the dimensionality reduction under non-negative constraints as above mentioned. Alternating least squares (ALS) [31] is

one of the most well known methods, and it can be computed by simple algorithm. Recently, independent component analysis (ICA), which is one of the multivariate analysis, was proposed [32, 33], and it has been mainly studied theoretically. ICA extracts independent components that are mutually statistically independent by using high order statistics (kurtosis). Non-negative ICA, which extracts independent components under non-negative constraints, was also proposed [34]. Non-negative matrix factorization (NMF) [35, 36, 37] was developed for image recognition, and its extension has been reported [38]. NMF has theoretical advantage to ALS in which the solution at least will reach local optimal solution. The curve resolution problem is still the challenging problem because it remains some practical problems such as the initial value that should be determined to compute these methods.

In part I, multivariate regression by PLS and regularized canonical correlation analysis (RCCA) was reported. RCCA has not been applied for the regression of high-dimensional data. We applied these methods to GC-MS data in which the intracellular metabolites of the leaves of Japanese green tea were analyzed. The main purpose of this research is to construct a quality-predictive model between ranking results by sensory test of tea taste and the value of chromatogram. The result shows that the optimal number of latent variables in RCCA was significantly fewer than in PLS, to construct a quality-predictive model.

In part II, visualization by multivariate analysis, PCA, PLS, OPLS, RFDA was achieved. We extended these methods to smoothness methods by introducing the differential penalty of the latent variables with each class, and to nonlinear method by kernel method. We applied these methods to CE-MS and GC-MS data in which the intracellular metabolites of yeast about ethanol fermentation from xylose were analyzed. The effect of smoothness can be found by the results.

In part III, curve resolution problem is achieved. We extended ALS to RALS (regularized ALS), an extension of ALS by using regularized term. (R)ALS and ICA

was applied to LC-diode array detector (DAD) data in which the part of the intracellular metabolites of microalgae was analyzed. The results suggested that RALS gives more optimal solution than ALS, and it gives more suitable solution with non-negativity than ICA.

References

[1] E. Fukusaki, A. Kobayashi, Plant metabolomics: potential for practical operation, J. Biosci. Bioeng. 100 (2005) 347–354.

[2] T. Soga, Y., Ohashi, Y. Ueno, H. Naraoka, M. Tomita, T. Nishioka, Quantitative metabolome analysis using capillary electrophoresis mass spectrometry, J. Proteome Res. 2 (2003) 488–494.

[3] K. Harada, E. Fukusaki, A. Kobayashi, Pressure-assisted capillary electrophoresis mass spectrometry using combination of polarity reversion and electroosmotic flow for metabolomics anion analysis, J Biosci Bioeng. 101 (2006) 403–409.

[4] T. Soga, R. Baran, M. Suematsu, Y. Ueno, S. Ikeda, T. Sakurakawa, Y. Kakazu, T. Ishikawa, M. Robert, T. Nishioka, M. Tomita, Differential metabolomics reveals ophthalmic acid as an oxidative stress biomarker indicating hepatic glutathione consumption, J Biol Chem. 16, 281 (2006) 16768–16776.

[5] R. Baran, H. Kochi, N. Saito, M. Suematsu, T. Soga, T. Nishioka, M. Robert, M. Tomita, MathDAMP: a package for differential analysis of metabolite profiles, BMC Bioinformatics, 7 (2006) 530–538.

[6] R.J. Bino, R.D. Hall, O. Fiehn, J. Kopka, K. Saito, J. Draper, B.J. Nikolau, P. Mendes, U. Roessner-Tunali, M.H. Beale, R.N. Trethewey, B.M. Lange, E.S. Wurtele, L.W. Sumner, Potential of metabolomics as a functional genomics tool, Trends Plant Sci., 9 (2004) 418–425.

[7] T. Tohge, Y. Nishiyama, M.Y. Hirai, M. Yano, J. Nakajima, M. Awazuhara, E. Inoue, H. Takahashi, D.B. Goodenowe, M. Kitayama, M. Noji, M. Yamazaki, K. Saito, Functional genomics by integrated analysis of metabolome and transcriptome of *Arabidopsis* plants over-expressing a MYB transcription factor, Plant J., 42 (2005) 218–235.

[8] L.M. Raamsdonk, B. Teusink, D. Broadhurst, N. Zhang, A. Hayes, M.C. Walsh, J.A. Berden, K.M. Brindle, D.B. Kell, J.J. Rowland, H.V. Westerhoff, K. van Dam,

S.G. Oliver, A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations, Nat. Biotechnol. 19 (2001) 45–50.

[9] J. Allen, H.M. Davey, D. Broadhurst, J.K. Heald, J.J. Rowland, S.G. Oliver, D.B. Kell, High-throughput classification of yeast mutants for functional genomics using metabolic footprinting, Nat Biotechnol. 21 (2003) 692–696.

[10] M.Y. Hirai, K. Sugiyama, Y. Sawada, T. Tohge, T. Obayashi, A. Suzuki, R. Araki, N. Sakurai, H. Suzuki, K. Aoki, H. Goda, O.I Nishizawa, D. Shibata, K. Saito, Omics-based identification of *Arabidopsis* Myb transcription factors regulating aliphatic glucosinolate biosynthesis, Proc. Natl. Acad. Sci., 104 (2007) 6478–6483.

[11] M.Y. Hirai, M. Yano, D.B. Goodenowe, S. Kanaya, T. Kimura, M., Awazuhara, M. Arita, T. Fujiwara, K. Saito, Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*, Proc. Natl. Acad. Sci., 101 (2004) 10205–10210.

[12] S. Kanaya, M. Kinouchi, T. Abe, Y. Kudo, Y. Yamada, T. Nishi, H. Mori, T. Ikemura, Analysis of codon usage diversity for bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome, Gene, 276 (2001) 89–99.

[13] M. Sugimoto, K. Shinichi, A. Masanori, S. Tomoyoshi, T. Nishioka, M. Tomita, Large-Scale prediction of cationic metabolite identity and migration time in capillary electrophoresis mass spectrometry using artificial neural networks, Anal. Chem., 77 (2005), 78–84.

[14] K. Shinoda, M. Sugimoto, N. Yachie, N. Sugiyama, T. Masuda, M. Robert, T. Soga, M. Tomita, Prediction of liquid chromatographic retention times of peptides generated by protease digestion of the *Escherichia coli* proteome using artificial neural networks, J Proteome Res., 5 (2006) 3312–3317.

[15] H. Martens, T. Naes, Multivariate calibration, John Wiley & Sons Ltd (1992).

[16] H. Akaike, A new look at the statistical model identification, IEEE trans. automat.

contr., 19 (1974) 716–723.

[17] B. Lavine, J. Workman, Chemometrics, Anal. Chem., 78 (2006) 4137–4145

[18] K. Fukunaga, Introduction to Statistical Pattern Recognition (2ed.), Academic Press, Inc., Boston, (1990).

[19] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intel. Lab. Syst., 58 (2001) 109–130.

[20] H. Xiaofei, P. Niyogi, Locality Preserving Projections, Advances in Neural Information Processing Systems 16, Vancouver, Canada, (2003).

[21] M. Sugiyama, Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis, JMLR, 8 (2007) 1027–1061.

[22] N. Krämer, A.L. Boulesteix, G. Tutz, Penalized partial least squares based on B-splines transformations, SFB 386, Discussion Paper 483 (2007)

[23] J. Ye, Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems, 6, (2005), 483–502.

[24] V. Vapnik, The Nature of Statistical Learning Theory. Springer-Verlag (1995).

[25] B. Schölkopf, A.J. Smola, K.R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comp. 10 (1998) 1299–1319.

[26] R. Rosipal, L.J. Trejo, Kernel partial least squares regression in reproducing kernel Hilbert space, JMLR 2 (2001) 97–123.

[27] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K.R. Müller, Fisher discriminant analysis with kernels, Neural Networks for Signal Processing IX (1999) 41–48.

[28] B. Schölkopf, K. Tsuda, J.P. Vert, Kernel Methods in Computational Biology, MIT Press, Cambridge (2004).

[29] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge (2004).

[30] B. Schölkopf, A.J. Smola, Learning With Kernels, MIT Press, Cambridge (2002).

[31] E.J. Karjalainen, The spectrum reconstruction problem, use of alternating regression for unexpected spectral components in two-dimensional spectroscopies,

Chemom. Intell. Lab. Syst. 7 (1989) 31–38.

[32] A. Hyvarinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Trans. Neural Netw., 10 (1999) 626–634.

[33] A. Hyvarinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley Interscience, New York (2001).

[34] M.D. Plumbley, Algorithms for non-negative independent component analysis, IEEE Trans. Neural Netw. 14 (2003) 534–543.

[35] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference, pp. 556–562, MIT Press (2001).

[36] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.

[37] R. Albright, J. Cox, D. Duling, A. Langville, C.D. Meyer, Algorithms, Initializations, and convergence for the nonnegative matrix factorization, NCSU Technical Report Math 81706 (2007).

[38] A. Cichocki, R. Zdunek, S. Amari, Csiszar's Divergences for non-negative matrix factorization: family of new algorithms, ICA2006, Charleston SC, USA, March 5-8, Springer LNCS 3889, (2006) 32–39.

SYNOPSIS

Part I
Multivariate analysis for regression

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

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), $N << p$. 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.

Part II
Multivariate analysis for visualization and discrimination

Dimensionality reduction by PCA, PLS, OPLS, RFDA with smoothed penalty

Dimensionality reduction is an important technique as a preprocessing of high-dimensional data. We extended some ordinary dimensionality reduction methods, principal component analysis (PCA), partial least squares (PLS), orthonormalized PLS, and regularized Fisher discriminant analysis (RFDA) by introducing the differential penalty of the latent variables with each class. We proposed smoothed PCA, PLS, OPLS, and RFDA for the data in which observation is in transition with time. A nonlinear extension to these methods by kernel methods was also proposed as kernel smoothed PCA, PLS, OPLS, and FDA. All these methods are formulated by generalized eigenvalue problem, so the solution can be computed easily. In this study, we applied these methods to the data in which the observation is in transition with time and the number of variables (p) exceeds the number of observation (N), $N << p$. In this paper, the effect of smoothness was elucidated by the results of visualization.

Part III
Multivariate analysis for curve resolution

Application of regularized alternating least squares and independent component analysis to curve resolution problem

The analysis of data from analytical equipment will be an important factor in the execution of metabolomics. Self-modeling curve resolution (SMCR) is one of the theoretical techniques of chemometrics and has recently been applied to the data of hyphenated chromatography techniques. Alternating least squares (ALS) is a classical SMCR method. In ALS, however, different solutions are produced depending on randomly chosen initial values. Simulation in the present study showed that the use of a normalized constraint in calculating ALS was effective in avoiding this problem. We also improved the ALS algorithm by adding a regularized term (regularized ALS: RALS). Independent component analysis (ICA) is a comparatively new method and has been discussed very actively by information science researchers, but has still been applied only in very few cases to curve resolution problems in chemometrics studies. We applied RALS with a normalized constraint and ICA to the HPLC-DAD data of *Haematococcus pluvialis* metabolites and obtained a high accuracy of peak detection, suggesting that these curve resolution methods are useful for identification of metabolites in metabolomics.

Part I

Multivariate analysis for regression

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

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 abovementioned 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] has 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), $N << p$, and where there is multicollinearity 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 $N << p$ 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 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 $N << p$ 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 **X** and the response variable **y** are mean-centered but are not scaled. Fifty-three samples were divided into two groups: forty-seven samples as a training set and six 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 min 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 **X** and the response variable **y**. However, MLR cannot be applied to $N << p$ type data. Regression methods by using latent variables such as PLS construct a regression model between a new explanatory variable **t**, which is obtained by dimensionality reduction of **X**, and the response variable **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)

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

$$\begin{aligned} &\max \quad [\mathrm{cov}(\mathbf{Xw}, \mathbf{y})]^2 \\ &\text{s.t.} \quad \mathbf{w}'\mathbf{w} = 1 \end{aligned}$$

where **w** is a weight vector. **X** and **y** are mean-entered. Finally, PLS is formulated as the following eigenvalue problem:

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

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 $\mathbf{t} = \mathbf{X}\mathbf{w}$. To calculate more than one latent variable, we perform deflation of **X** and **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)

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

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

This conditional equation is rewritten as follows:

$$\max \quad [\mathrm{cov}(\mathbf{X}\mathbf{w},\mathbf{y})]^2$$
$$\text{s.t.} \quad \frac{1}{N}\mathbf{w}'\mathbf{X}'\mathbf{X}\mathbf{w} = 1$$

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

$$\frac{1}{N}\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}\mathbf{w} = \lambda\mathbf{X}'\mathbf{X}\mathbf{w} \qquad (2)$$

This generalized eigenvalue problem is solved by Cholesky decomposition of **X'X** when **X'X** is full rank and SVD.

2.2.3. Regularized canonical correlation analysis (RCCA)

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

$$\begin{aligned} &\max \quad [\mathrm{cov}(\mathbf{Xw}, \mathbf{y})]^2 \\ &\text{s.t.} \quad (1-\tau)\frac{1}{N}\mathbf{w'X'Xw} + \tau \mathbf{w'w} = 1 \end{aligned}$$

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

$$\frac{1}{N^2}\mathbf{X'yy'Xw} = \lambda\left\{(1-\tau)\frac{1}{N}\mathbf{X'X} + \tau \mathbf{I}\right\}\mathbf{w} \qquad (3)$$

To calculate more than one latent variable, we perform deflation of **X** and **y** and calculate the eigenvector corresponding to the maximum eigenvalue in Eq. (3) as with PLS. This generalized eigenvalue problem is solved by Cholesky decomposition and SVD. In Eq. (3), $\tau = 1$ corresponds to PLS while $\tau = 0$ corresponds to CCA.

2.2.4. Kernel PLS

Kernel methods start by mapping the original data onto a high-dimensional feature space corresponding to the reproducing kernel Hilbert space (RKHS). The inner product of samples φ_x and φ_y in the feature space $<\varphi_x, \varphi_y>$ can be replaced by the kernel function $k(\mathbf{x}, \mathbf{y})$ which is positive definite:

$$\left\langle \varphi_x, \varphi_y \right\rangle = k(\mathbf{x}, \mathbf{y})$$

This formulation is called the "kernel trick". We can extend multivariate analysis to nonlinear multivariate analysis using the kernel trick without explicitly knowing the

mapping in the feature space. As kernel functions, polynomial kernel and Gaussian kernel have been widely used. A detailed theory of kernel methods has been outlined previously [12, 13]

Rosipal et al. [17] introduced kernel PLS using a nonlinear iterative partial least squares (NIPALS) algorithm. Similarly to PLS, kernel PLS is explained as the optimization problem of maximizing the square of covariance between the score vector **t**, which is a linear combination of the explanatory variable Φ in the feature space, and the response variable **y**. We assume that there exists a coefficient α which satisfies $\mathbf{w} = \Phi'\alpha$ and denotes $\mathbf{K} = \Phi\Phi'$:

$$\begin{aligned} &\max \quad [\mathrm{cov}(\mathbf{K}\boldsymbol{\alpha}, \mathbf{y})]^2 \\ &\text{s.t.} \quad \frac{1}{N}\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} = 1 \end{aligned}$$

Finally, kernel PLS is formulated as the following generalized eigenvalue problem:

$$\frac{1}{N^2}\mathbf{K}\mathbf{y}\mathbf{y}'\mathbf{K}\boldsymbol{\alpha} = \lambda\mathbf{K}\boldsymbol{\alpha} \qquad (4)$$

K is a mean-centered kernel gram matrix of which the elements are an output of the kernel function. We perform deflation to **K** and **y** and then calculate the eigenvector corresponding to the maximum eigenvalue in Eq. (4) to calculate more than one latent variable. When **K** is rank-deficient, we approximate **K** on the righthand side of Eq. (4) to $\mathbf{K} + \varepsilon\,\mathbf{I}$ where ε is a small positive value. This generalized eigenvalue problem is solved by Cholesky decomposition and SVD. A score vector can be calculated as $\mathbf{t} = \mathbf{K}\alpha$.

2.2.5. Kernel CCA

Akaho [18] and Bach and Jordan [19] introduced kernel CCA, which is a kernelized version of RCCA. Like kernel PLS, we assume that there exists a coefficient

α which satisfies $\mathbf{w} = \Phi'\alpha$ and denotes $\mathbf{K} = \Phi\Phi'$:

$$\begin{aligned} &\max \quad [\mathrm{cov}(\mathbf{K}\boldsymbol{\alpha}, \mathbf{y})]^2 \\ &\text{s.t.} \quad (1-\tau)\frac{1}{N}\boldsymbol{\alpha}'\mathbf{K}^2\boldsymbol{\alpha} + \tau\,\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} = 1 \end{aligned}$$

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

$$\frac{1}{N^2}\mathbf{K}\mathbf{y}\mathbf{y}'\mathbf{K}\boldsymbol{\alpha} = \lambda\left\{(1-\tau)\frac{1}{N}\mathbf{K}^2 + \tau\,\mathbf{K}\right\}\boldsymbol{\alpha} \qquad (5)$$

Bach and Jordan [19] approximated this eigenvalue problem for computational simplicity as follows:

$$\frac{1}{N^2}\mathbf{K}\mathbf{y}\mathbf{y}'\mathbf{K}\boldsymbol{\alpha} = \lambda\left(\mathbf{K} + \frac{N\kappa}{2}\mathbf{I}\right)^2\boldsymbol{\alpha}$$

where κ is $\tau/(1-\tau)$.

Let κ be a small positive value. We perform deflation of $\mathbf{K}$ and $\mathbf{y}$ and then calculate the eigenvector corresponding to the maximum eigenvalue in Eq. (5) to calculate more than one latent variable, as in kernel PLS. This generalized eigenvalue problem is solved by Cholesky decomposition and SVD. A score vector can be calculated as $\mathbf{t} = \mathbf{K}\alpha$.

The algorithm for multivariate regressions is given in the Appendix. Table 1 gives a summary of these methods.

2.3. Software

The computer program for PLS, RCCA, kernel PLS, and kernel CCA calculation was developed in-house by the authors with the personal computer version of MATLAB 7 (Mathworks, Natick, MA, USA) by the authors. The computer system used to run these programs has a 2.4 GHz CPU and a 512 MB memory.

Table 1 Summary of partial least squares (PLS), regularized canonical correlation analysis (RCCA), kernel PLS, and kernel CCA where the response variable is univariate

	PLS	RCCA	Kernel PLS	Kernel CCA
w or α	$\mathbf{w} = \mathbf{X}'\mathbf{y}$	$\mathbf{w} = \left\{(1-\tau)\frac{1}{N}\mathbf{X}'\mathbf{X} + \tau\mathbf{I}\right\}^{-1}\mathbf{X}'\mathbf{y}$	$\boldsymbol{\alpha} = \mathbf{y}$	$\boldsymbol{\alpha} = \left\{(1-\tau)\frac{1}{N}\mathbf{K} + \tau\mathbf{I}\right\}^{-1}\mathbf{y}$
	$\mathbf{w} \leftarrow \frac{\mathbf{w}}{\sqrt{\mathbf{w}'\mathbf{w}}}$	$\mathbf{w} \leftarrow \frac{\mathbf{w}}{\sqrt{(1-\tau)\frac{1}{N}\mathbf{w}'\mathbf{X}'\mathbf{X}\mathbf{w} + \tau\mathbf{w}'\mathbf{w}}}$	$\boldsymbol{\alpha} \leftarrow \frac{\boldsymbol{\alpha}}{\sqrt{\frac{1}{N}\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}}}$	$\boldsymbol{\alpha} \leftarrow \frac{\boldsymbol{\alpha}}{\sqrt{(1-\tau)\frac{1}{N}\boldsymbol{\alpha}'\mathbf{K}^2\boldsymbol{\alpha} + \tau\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}}}$

3. Results and discussion

3.1. Construction of a quality-predictive model by PLS and RCCA

For the training set, we determined the optimal number of latent variables of PLS, RCCA (τ = 0.5), and RCCA (τ = 0.1) using leave-one-out cross validation (LOOCV). The optimal number of latent variables was 15, 3, and 1, respectively (Fig. 1). The root mean squared error (RMSE) of PLS and RCCA (τ = 0.1) was 3.0054 and 3.6631 for the training set and 12.7368 and 10.4296 for the test set, respectively (Table 2). These results indicate that the number of components required to construct a quality-predictive model in RCCA is fewer than in PLS. The reason is thought to be as explained below.

Because PLS maximizes the covariance between the score vector **t**, which is a linear combination of **X**, and the response variable **y**, the PLS model is affected by the explanatory variable **X**, which is not closely related to the response variable **y** and has a

Table 2 Number of latent variables (LV) and root mean squared error (RMSE) by PLS and RCCA (τ = 0.5 or 0.1)

	Number of LV	RMSE	
		Training set	Test set
PLS	15	3.0054	12.7368
RCCA (τ = 0.5)	3	3.4639	11.6177
RCCA (τ = 0.1)	1	3.6631	10.4296

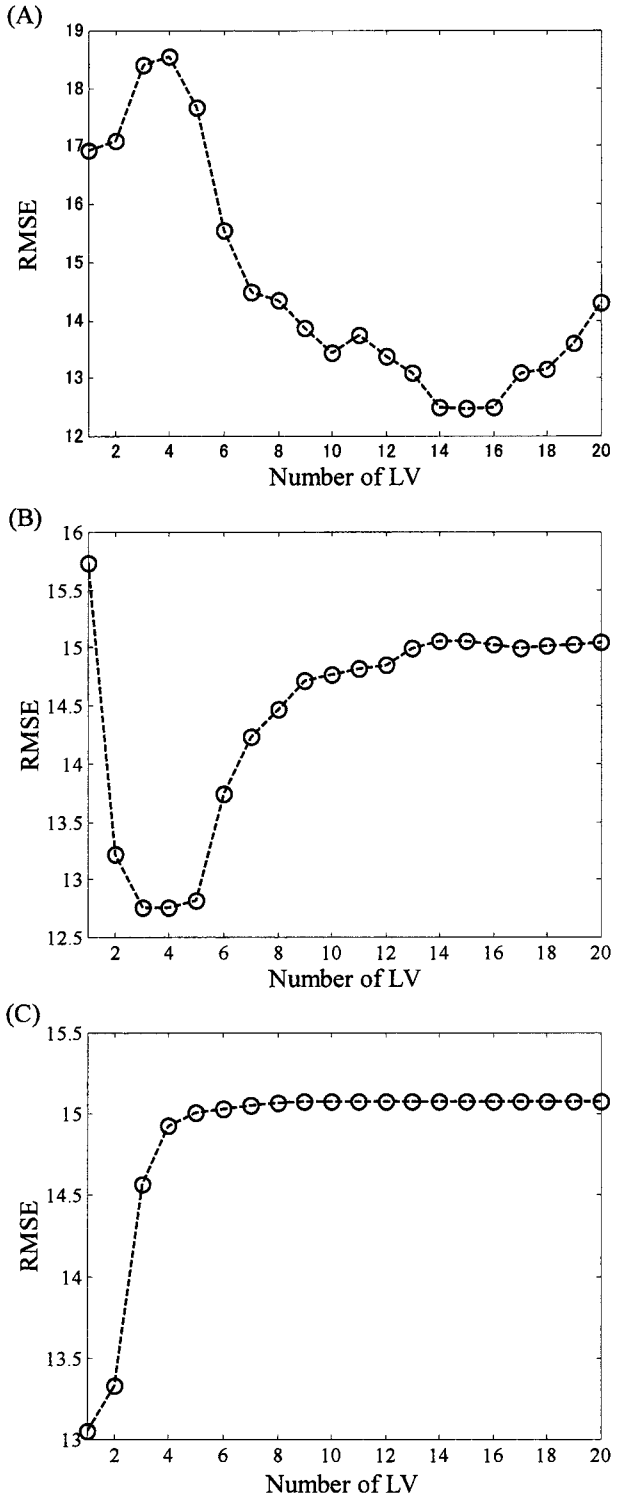

Fig. 1. Leave-one-out cross-validation (LOOCV) of PLS (A) and RCCA ($\tau = 0.5$ (B), $\tau = 0.1$ (C)).

large variance. As a result, PLS extracts a number of latent variables that are redundant to the construction of a quality-predictive model. In the practical application of PLS, the explanatory variables, which are respective columns of **X**, are often scaled to unit variance to reduce this effect. In contrast, CCA maximizes the correlation between the score vector **t** and the response variable **y**. Because the correlation coefficient is equal to the covariance, which is divided by the product of the standard deviations of **t** and **y**, the effect of the explanatory variable **X**, which is not closely related to the response variable **y** and has a large variance, is reduced in RCCA.

3.2. Reduction of calculation time of RCCA

A long calculation time is required for cross-validation of RCCA. The calculation time for LOOCV of RCCA by implementation of Eq. (3) was 5.9962×10^4 s (about 16.5 h), while that for LOOCV of PLS was 54.2970 s. By using kernel CCA with a linear kernel, we were able to reduce the calculation time for LOOCV of RCCA to 8.1090 s. The main reason for this reduction is that the former formulation of RCCA required the handling of a large matrix of the size 2064×2064; whereas the latter formulation requires only a small size matrix of the size 47×47. A similar idea to reduce the calculation time and memory demand for PLS was proposed by Lindgren et al. [20].

3.3. Metabolic fingerprinting with PLS and RCCA

For metabolic fingerprinting, variable importance in projection (VIP) has often been used. From the VIP value, we can identify metabolites important for the quality-predictive model by reading the retention time of the chromatogram and its mass spectrum. We picked out ten metabolites in descending order of VIP value as shown in Table 3.

Boulesteix et al. [7] pointed out that the VIP index lacks a theoretical background. In the present study, we used only one component for the calculation of the VIP value. In this case, the VIP value is equal to the root of the square of the weight vector multiplied by the number of variables. Moreover, the weight vector of PLS is proportional to the covariance between the explanatory variable **X** and the response variable **y** [23]. Therefore, the VIP value of PLS in the present study corresponds to the covariance between the explanatory variable **X** and the response variable **y**. Table 1 indicates that the weight vector and VIP value of RCCA correspond to the regression coefficient of ridge regression.

Table 3 indicates that the result of PLS was very similar to that of the orthogonal signal correction PLS (OSC-PLS) reported in a previous study [21], except for theanine. Quinic acid, amino acids, and groups of sugars were found to be significant in the construction of the quality-prediction model described in the previous study [21]. Furthermore, the VIP values corresponding to sucrose and glucose were significantly smaller in RCCA than in PLS and OSC-PLS. In contrast, the VIP values corresponding to glutamic acid and oxalic acid were larger in RCCA than in PLS and OSC-PLS, suggesting that these metabolites may also have a considerable effect on the quality-prediction model.

Table 3 VIP (variable importance in projection) value and ten metabolites identified by PLS and RCCA

Ranking	VIP value		Metabolite		
	PLS	RCCA ($\tau = 0.1$)	PLS	RCCA ($\tau = 0.1$)	OSC-PLS*
1	23.8273	8.0146	sucrose	glutamic acid	sucrose
2	13.8409	7.9499	glucose	fructose1	glucose
3	8.5963	7.3738	malic acid	quinic acid	quinic acid
4	8.5884	7.2561	phosphoric acid	caffeine	fructose1
5	8.3986	7.1	fructose1	oxalic acid	caffeine
6	8.0194	6.9783	glutamine2	theanine	malic acid
7	7.2698	6.2757	quinic acid	malic acid	theanine
8	5.8229	6.2279	caffeine	silane	fructose2
9	5.2665	6.2035	fructose2	glucose	glutamine2
10	5.1425	6.0373	silane	inositol	phosphoric acid

* Orthogonal signal correction PLS [21]

3.4. Kernel PLS and kernel CCA for multivariate regression

As in linear PLS, the explanatory variable Φ in the feature space, which is not closely related to the response variable **y** and has a large variance, affects the quality-prediction model using kernel PLS. The effect may be more serious in kernel PLS than in linear PLS because we cannot perform scaling for the explanatory variable Φ directly in the feature space.

We also applied kernel PLS and kernel CCA to the construction of a multivariate regression model (data not shown). As long as we searched for the optimal parameter of the kernel function, the quality-prediction model constructed with kernel PLS and kernel CCA was not significantly better than that with PLS and RCCA. We therefore concluded that linear PLS and linear RCCA are adequate for the construction of a quality-prediction model. In cases where the nonlinear relation between **X** and **y** is strong, the kernel method will be useful in constructing a multivariate regression model.

Application of the kernel method to multivariate regression has been studied [24, 25], but there have been few applications. Kernel PLS and kernel CCA for multivariate regression need to be applied to various datasets and the usefulness of these methods demonstrated.

Appendix

A.1. Algorithm for PLS, RCCA, kernel PLS, and kernel CCA

The pseudo-code of the algorithms for multivariate regression is as follows:

For $i = 1$ to k

Compute **w** or α shown in Table1 / Solve generalized eigenvalue problem.

Normalize **w** or α according to the constraint.

$\mathbf{t} = \mathbf{Xw}$ or $\mathbf{t} = \mathbf{K}\alpha$.

Deflation of **X** or **K**, and **y**.

End

We perform regression between **T** and **y**, $\mathbf{y} = \mathbf{Tb}$. The regression coefficient **b** is written as follows:

$$\mathbf{b} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}$$

The matrix **T** has a score vector **t** in each column, $T=[t_1\ t_2\ \cdots\ t_k]$. For the training set, the prediction is written as follows:

$$\hat{\mathbf{y}} = \mathbf{Tb} = \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}$$

For the test set, the prediction is written as follows:

$$\hat{\mathbf{y}}_t = \mathbf{T}_t\mathbf{b} = \mathbf{T}_t(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}$$

A.2. Deflation for multivariate regression

The data matrix **X** is approximated as follows:

$$\mathbf{X} \approx \mathbf{t}_1\mathbf{p}_1 + \mathbf{t}_2\mathbf{p}_2 + \cdots + \mathbf{t}_k\mathbf{p}_k \qquad \text{(A1)}$$

where **p** is called a loading vector. We assume that the first term in Eq. (A1), $\mathbf{t}_1\mathbf{p}_1$, is much larger than the sum of the other terms, $\mathbf{t}_2\mathbf{p}_2 + \cdots + \mathbf{t}_k\mathbf{p}_k$. Then, the loading vector **p** is approximated as follows:

$$\mathbf{p}_1 = \frac{\mathbf{t}_1'\mathbf{X}}{\mathbf{t}_1'\mathbf{t}_1}$$

Deflation of **X** and similarly of **y** is performed as follows:

$$\mathbf{X} \leftarrow \mathbf{X} - \mathbf{t}\frac{\mathbf{t}'\mathbf{X}}{\mathbf{t}'\mathbf{t}}$$

$$\mathbf{y} \leftarrow \mathbf{y} - \mathbf{t}\frac{\mathbf{t}'\mathbf{y}}{\mathbf{t}'\mathbf{t}}$$

Deflation of Φ in the feature space is performed by deflation of **K** as follows:

$$\mathbf{K} = \left(\mathbf{\Phi} - \mathbf{t}_1 \frac{\mathbf{t}_1'\mathbf{\Phi}}{\mathbf{t}_1'\mathbf{t}_1}\right)\left(\mathbf{\Phi} - \mathbf{t}_1 \frac{\mathbf{t}_1'\mathbf{\Phi}}{\mathbf{t}_1'\mathbf{t}_1}\right)' = \left(\mathbf{I} - \mathbf{t}_1 \frac{\mathbf{t}_1'}{\mathbf{t}_1'\mathbf{t}_1}\right)\mathbf{K}\left(\mathbf{I} - \mathbf{t}_1 \frac{\mathbf{t}_1'}{\mathbf{t}_1'\mathbf{t}_1}\right)'$$

By deflation, we obtain mutually orthogonal score vectors.

References

[1] E. Fukusaki, A. Kobayashi, Plant metabolomics: potential for practical operation, J. Biosci. Bioeng. 100 (2005) 347–354.

[2] O. Fiehn, Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks, Comp. Funct. Genom. 2 (2001) 155–168.

[3] L.M. Raamsdonk, B. Teusink, D. Broadhurst, N. Zhang, A. Hayes, M.C. Walsh, J.A. Berden, K.M. Brindle, D.B. Kell, J.J. Rowland, H.V. Westerhoff, K. van Dam, S.G. Oliver, A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations, Nat. Biotechnol. 19 (2001) 45–50.

[4] M.R. Antoniewicz, G. Stephanopoulos, J.K. Kelleher, Evaluation of regression models in metabolic physiology: predicting fluxes from isotopic data without knowledge of the pathway, Metabolomics 2 (2006) 41–52.

[5] J. Gullberga, P. Jonssonb, A. Nordströma, M. Sjöströmb, T. Moritz, Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of Arabidopsis thaliana samples in metabolomic studies with gas chromatography/mass spectrometry, Anal. Biochem. 331 (2004) 283–295.

[6] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics. Chemometr. Intel. Lab. Syst. 58 (2001) 109–130.

[7] A. Boulesteix, K. Strimmer, Partial least squares: a versatile tool for the analysis of high-dimensional genomic data, Brief Bioinform. 8 (2007) 32–44.

[8] D. Nguyen, D. Rocke, Tumor classification by partial least squares using microarray gene expression data, Bioinformatics. 18 (2002) 39–50.

[9] A. Boulesteix, PLS dimension reduction for classification with microarray data, Stat. Appl. Genet. & Mol. Biol. 3 (2004) Article 33.

[10] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, Neural Comp. 16 (2004) 2639–2664.

[11] R. Rosipal, N. Krämer, Overview and recent advances in partial least squares, in: C. Saunders, M. Grobelnik, S. Gunn, J. Shawe-Taylor (Eds.), Subspace, Latent Structure and Feature Selection Techniques, Springer, New York, 2006, pp. 34–51.

[12] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, 2004.

[13] B. Schölkopf, A.J. Smola, Learning With Kernels, MIT Press, Cambridge, 2002.

[14] B. Schölkopf, K. Tsuda, J.P. Vert, Kernel Methods in Computational Biology, MIT Press, Cambridge, 2004.

[15] B. Schölkopf, A.J. Smola, K.R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comp. 10 (1998) 1299–1319.

[16] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K.R. Müller, Fisher discriminant analysis with kernels, Neural Networks for Signal Processing IX (1999) 41–48.

[17] R. Rosipal, L.J. Trejo, Kernel partial least squares regression in reproducing kernel Hilbert space, JMLR 2 (2001) 97–123.

[18] S. Akaho, A kernel method for canonical correlation analysis, International Meeting of Psychometric Society (2001).

[19] F.R. Bach, M.I. Jordan, Kernel independent component analysis, JMLR 3 (2002) 1–48.

[20] F. Lindgren, P. Geladi, S. Wold, The kernel algorithm for PLS, J. Chemometr. 7 (1993) 45–59.

[21] W. Pongsuwan, E. Fukusaki, T. Bamba, T. Yonetani, T. Yamahara, A. Kobayashi, Prediction of Japanese green tea ranking by gas chromatography/mass spectrometry-based hydrophilic metabolite fingerprinting, J Agric. Food Chem. 55 (2007) 231–236.

[22] M. Borga, T. Landelius, H, Knutsson, A unified approach to PCA, PLS, MLR and CCA, Technical Report, Linköping University, 1997.

[23] J.A. Wegelin, A survey of partial least squares (PLS) methods, with emphasis on

the two-block case, Technical Report, University of Washington, 2000.

[24] R.P. Cogdilla, P. Dardenneb, Least-squares support vector machines for chemometrics: an introduction and evaluation, J. Near Infrared Spectrosc. 12 (2004) 93–100.

[25] H. Shinzawa, J.H. Jiang, P. Ritthiruangdej, Y. Ozaki, Investigations of bagged kernel partial least squares (KPLS) and boosting KPLS with applications to near-infrared (NIR) spectra, J. Chemometr. 20 (2006) 436–444.

Part II

Multivariate analysis for visualization and discrimination

Dimensionality reduction by PCA, PLS, OPLS, RFDA with smoothness

1. Introduction

Recently, some comprehensive analysis such as gene expression analysis from microarray and metabolome analysis has been studied actively. These data are often high-dimensional data, especially $N << p$, the number of variables (p) exceeds the number of observation (N), type data. Dimensionality reduction technique is required to visualize high-dimensional data as a preprocessing step. An unsupervised dimensionality reduction methods such as principal component analysis (PCA) and a supervised dimensionality reduction such as partial least squares (PLS) for discrimination [1] and Fisher discriminant analysis (FDA) [2] have been applied.

For time series data analysis, smoothed PCA [3, 4] was proposed in the context of functional data analysis [5]. The differential penalty of the weight vector is added to the constraint condition of PCA in smoothed PCA, so it is suitable for the data in which the variables are in transition with time. In this study, we deal with the data in which the observation is in transition with time, and the data of such a type sometimes appears in microarray and metabolome research. In this study, we extended some ordinary dimensionality reduction methods, PCA, PLS, orthonormalized PLS (OPLS) [6], regularized FDA (RFDA) [7] by introducing the differential penalty with each class. We proposed smoothed PCA, PLS, PLS, RFDA and its nonlinear extension by kernel method as kernel smoothed PCA, PLS, OPLS, and FDA. These methods can reflect information about the data in which the observation is in transition with time to the results of visualization.

In the present study, we applied smoothed PCA, PLS, OPLS, and RFDA, and kernel smoothed PCA, PLS, OPLS, FDA to CE-MS and GC-MS data, which is high-dimensional data ($N << p$). The main objective of the present study is to evaluate

these newly methods for metabolome data. This metabolome data is collected to clarify the part of the phenomenon of intracellular xylose metabolism in yeast for the fermentative production of ethanol by metabolome analysis.

2. Theory

We explained about smoothness by using differential penalty, and dimensionality reduction methods such as PCA, PLS, OPLS, and RFDA, and its extension by introducing differential penalty and its nonlinear extension by kernel methods. These methods are formulated by generalized eigenvalue problem.

2.1. Differential penalty

Eilers [8] introduced the smoothing method by using differential penalty as 'Whittaker smoother'. This problem is formulated as minimization problem as follows:

$$\min \quad \|\mathbf{s}-\mathbf{t}\|^2 + \kappa\|\mathbf{Ds}\|^2 \qquad (1)$$

The observation series **t** is fitted to the smoothed series **s**. First and second differential matrix D is set as follows in the case, which the number of observations is 5.

$$\mathbf{D'}=\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \quad \mathbf{D''}=\begin{bmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{bmatrix}$$

κ is parameter of smoothness. The first term in Eq. (1) is the squared error between **s** and **t** and the second term is differential penalty of smooth series **s**. As the parameter of smoothness κ becomes smaller, the squared error between **s** and **t** becomes more important. As κ becomes larger, the smoothness of **s** becomes strong. The minimization of Eq. (1) is trade-off between the squared error and smoothness. The smoothness is controlled by the parameter of smoothness κ.

2.2. Linear multivariate analysis with smoothness

In the previous study, smoothed PCA [3, 4] was introduced in the differential penalty of the weight vectors in the context of functional data analysis [5] and penalized PLS was also proposed [9]. The differential penalty in the context of functional data analysis was interpreted as the smoothness that is suited for the data in which the variables were in transition with time. In this study, we introduced the differential penalty of the latent variables with each class to PCA, PLS, OPLS, and RFDA, and these methods are suited for the data in which the observation is in transition with time. The differential matrix D is set for the g class classification problem as follows:

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_g \end{bmatrix}$$

2.2.1. Principal component analysis (PCA)

PCA is explained as the optimization problem of maximizing the variance of the latent variables **t**, which is the linear combination of the explanatory variable **X**.

$$\begin{aligned} &\max \quad \mathrm{var}(\mathbf{t}) \\ &\text{s.t.} \quad \mathbf{w}'\mathbf{w} = 1 \end{aligned}$$

where **w** is a weight vector. **X** is mean-centered. PCA is written as the following eigenvalue problem:

$$(1/N)\mathbf{X}'\mathbf{X}\mathbf{w} = \lambda\mathbf{w}$$

We added the differential penalty of the latent variables **t** to the constraint condition of PCA. Smoothed PCA is formulated as follows:

$$\begin{aligned} &\max \quad \mathrm{var}(\mathbf{t}) \\ &\text{s.t.} \quad \mathbf{w}'\mathbf{w} + \kappa(\mathbf{D}\mathbf{t})'(\mathbf{D}\mathbf{t}) = 1 \end{aligned}$$

Finally, smoothed PCA is written as the following generalized eigenvalue problem about the weight vector **w**:

$$\frac{1}{N}\mathbf{X'Xw} = \lambda(\mathbf{I} + \kappa\mathbf{X'D'DX})\mathbf{w}$$

$\mathbf{w}_i$ is weight vector corresponds to the large eigenvalue in i-th. The i-th latent variables $\mathbf{t}_i$ can be calculated as $\mathbf{t}_i=\mathbf{X}\mathbf{w}_i$.

2.2.2. Partial least squares (PLS)

PLS is explained as the optimization problem of maximizing the covariance between the latent variables $\mathbf{t}=\mathbf{X}\mathbf{w}_x$, which is a linear combination of the explanatory variable **X**, and the latent variables $\mathbf{s}=\mathbf{Y}\mathbf{w}_y$, which is a linear combination of the response variable **Y**.

$$\begin{aligned} &\max \quad \mathrm{cov}(\mathbf{t},\mathbf{s}) \\ &\text{s.t.} \quad \mathbf{w}_x'\mathbf{w}_x = 1 \quad \mathbf{w}_y'\mathbf{w}_y = 1 \end{aligned}$$

X and **Y** are mean-centered. Finally, PLS is written as the following eigenvalue problem about the weight vector $\mathbf{w}_x$:

$$\frac{1}{N}\mathbf{X'YY'X}\mathbf{w}_x = \lambda\mathbf{w}_x \qquad (2)$$

We added the differential penalty of the latent variables **t** to the constraint condition of PLS. Smoothed PLS is formulated as follows:

$$\begin{aligned} &\max \quad \mathrm{cov}(\mathbf{t},\mathbf{s}) \\ &\text{s.t.} \quad \mathbf{w}_x'\mathbf{w}_x + \kappa(\mathbf{Dt})'(\mathbf{Dt}) = 1 \quad \mathbf{w}_y'\mathbf{w}_y = 1 \end{aligned}$$

Finally, smoothed PLS is written as the following generalized eigenvalue problem about the weight vector $\mathbf{w}_x$:

$$\frac{1}{N^2}\mathbf{X'YY'X}\mathbf{w}_x = \lambda(\mathbf{I} + \kappa\mathbf{X'D'DX})\mathbf{w}_x$$

The matrix **I** is identity matrix.

2.2.3. Orthonormalized partial least squares (OPLS)

OPLS is explained as the optimization problem of maximizing the covariance between the latent variables **t** and **s** as well as PLS. The constraint condition of OPLS is different from PLS.

$$\begin{aligned} &\max \quad \mathrm{cov}(\mathbf{t},\mathbf{s}) \\ &\text{s.t.} \quad \mathbf{w}_x'\mathbf{w}_x = 1 \quad \mathrm{var}(\mathbf{s}) = 1 \end{aligned}$$

Finally, OPLS is written as the following eigenvalue problem about the weight vector $\mathbf{w}_X$:

$$\frac{1}{N}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}\mathbf{w}_x = \lambda\mathbf{w}_x \qquad (3)$$

We added the differential penalty of the latent variables **t** to the constraint condition of OPLS. Smoothed OPLS is formulated as follows:

$$\begin{aligned} &\max \quad \mathrm{cov}(\mathbf{t},\mathbf{s}) \\ &\text{s.t.} \quad \mathbf{w}_x'\mathbf{w}_x + \kappa(\mathbf{D}\mathbf{t})'(\mathbf{D}\mathbf{t}) = 1 \quad \mathrm{var}(\mathbf{s}) = 1 \end{aligned}$$

Finally, smoothed OPLS is written as the following generalized eigenvalue problem about the weight vector $\mathbf{w}_X$:

$$\frac{1}{N^2}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}\mathbf{w}_x = \lambda(\mathbf{I} + \kappa\mathbf{X}'\mathbf{D}'\mathbf{D}\mathbf{X})\mathbf{w}_x$$

2.2.4. Regularized Fisher discriminant analysis

FDA is equal to canonical correlation analysis (CCA) when the response variables are a certain formulation of dummy variables [1]. We explained FDA by using the formulation of CCA. FDA is explained as the optimization problem of maximizing the correlation between the latent variables **t** and **s**.

$$\max \quad corr(\mathbf{t},\mathbf{s})$$

This conditional equation is rewritten as follows:

$$\begin{aligned} &\max \quad \mathrm{cov}(\mathbf{t},\mathbf{s}) \\ &\text{s.t.} \quad \mathrm{var}(\mathbf{t}) = 1 \quad \mathrm{var}(\mathbf{s}) = 1 \end{aligned}$$

However, FDA is not applicable to the $N << p$ type data for $\mathbf{X}$ or $\mathbf{Y}$. Here we considered the case in which the explanatory variables $\mathbf{X}$ are $N << p$, so l_2-norm regularized term of the weight vector $\mathbf{w}_x$ is introduced to FDA. FDA with regularized term, regularized FDA (RFDA), is formulated as follows:

$$\begin{aligned} &\max \quad \mathrm{cov}(\mathbf{t},\mathbf{s}) \\ &\text{s.t.} \quad (1-\tau)\,\mathrm{var}(\mathbf{t}) + \tau \mathbf{w}_x'\mathbf{w}_x = 1 \quad \mathrm{var}(\mathbf{s}) = 1 \end{aligned}$$

τ is regularized parameter and $0 \leqq \tau \leqq 1$. Finally, RFDA is written as the following generalized eigenvalue problem about the weight vector $\mathbf{w}_x$:

$$\frac{1}{N}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}\mathbf{w}_x = \lambda\left\{(1-\tau)\frac{1}{N}\mathbf{X}'\mathbf{X} + \tau\mathbf{I}\right\}\mathbf{w}_x$$

We added the differential penalty of the latent variables $\mathbf{t}$ to the constraint condition of RFDA. Smoothed RFDA is formulated as follows:

$$\begin{aligned} &\max \quad \mathrm{cov}(\mathbf{t},\mathbf{s}) \\ &\text{s.t.} \quad (1-\tau)\,\mathrm{var}(\mathbf{t}) + \tau \mathbf{w}_x'\mathbf{w}_x + \kappa(\mathbf{D}\mathbf{t})'(\mathbf{D}\mathbf{t}) = 1 \quad \mathrm{var}(\mathbf{s}) = 1 \end{aligned}$$

Finally, smoothed RFDA is written as the following generalized eigenvalue problem about the weight vector $\mathbf{w}_x$:

$$\frac{1}{N^2}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}\mathbf{w}_x = \lambda\left\{(1-\tau)\frac{1}{N}\mathbf{X}'\mathbf{X} + \tau\mathbf{I} + \kappa\mathbf{X}'\mathbf{D}'\mathbf{D}\mathbf{X}\right\}\mathbf{w}_x$$

2.3. How to set dummy variables to response variables for supervised dimensionality reduction

In discrimination by PLS, OPLS, RFDA, the response variable is set dummy variables as follows:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{1}_{n1} & \mathbf{0}_{n1} & \cdots & \mathbf{0}_{n1} \\ \mathbf{0}_{n2} & \mathbf{1}_{n2} & \cdots & \mathbf{0}_{n2} \\ \vdots & \vdots & \ddots & \mathbf{1}_{ng-1} \\ \mathbf{0}_{ng} & \mathbf{0}_{ng} & \cdots & \mathbf{0}_{ng} \end{bmatrix}$$

Barker et al. [1] suggested that FDA is equal to CCA when the response variables are dummy variables, and they defined OPLS as PLS for discrimination. OPLS is explained as the special case of FDA that only maximizes the scatter among class. The difference between PLS and OPLS is prior probabilities of class. The prior probabilities of PLS is the squared of n_i/N, and OPLS is n_i/N. The n_i is the number of observations within i-th class. Recently, the prior probability of PLS was reported in detail [10].

2.4. Nonlinear multivariate analysis by kernel method with smoothness

We proposed kernel smoothed PCA, PLS, OPLS, and FDA, which is the nonlinear extension of smoothed PCA, PLS, OPLS, RFDA by kernel methods. Kernel methods start by mapping the original data onto a high-dimensional feature space corresponding to the reproducing kernel Hilbert space (RKHS). The inner product of samples φ_x and φ_y in the feature space $\langle\varphi_x, \varphi_y\rangle$ can be replaced by the kernel function $k(\mathbf{x}, \mathbf{y})$ which is positive definite:

$$\langle \varphi_x, \varphi_y \rangle = k(\mathbf{x}, \mathbf{y})$$

This formulation is called the "kernel trick". We can extend multivariate analysis to nonlinear multivariate analysis using the kernel trick without explicitly knowing the mapping in the feature space. As kernel functions, polynomial kernel and Gaussian kernel have been widely used. A detailed theory of kernel methods has been outlined previously [7]. The advantageous point of the formulation by kernel methods is that we can compute nonlinear multivariate analysis easily and the computational costs for $N << p$ data, especially leave-one-out cross validation, is reduced by using linear kernel [11].

2.4.1. Kernel PCA

Kernel PCA [12] is explained as the optimization problem of maximizing the variance of the latent variables $\mathbf{t}= \Phi \mathbf{w}$, which is the linear combination of the explanatory variable Φ in feature space. We assume that there exists a coefficient α which satisfies $\mathbf{w} = \Phi'\alpha$ and denotes $\mathbf{K} = \Phi\Phi'$. The latent variables $\mathbf{t}$ can be calculated as $\mathbf{t}=\mathbf{K}\alpha$.

$$\begin{aligned} &\max \quad \mathrm{var}(\mathbf{t}) \\ &\text{s.t.} \quad \boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} = 1 \end{aligned}$$

$\mathbf{K}$ is mean-centered. Finally, kernel PCA is written as the following generalized eigenvalue problem about the coefficients vector α:

$$\frac{1}{N}\mathbf{K}^2\boldsymbol{\alpha} = \lambda\mathbf{K}\boldsymbol{\alpha}$$

When $\mathbf{K}$ is rank-deficient, we approximate $\mathbf{K}$ on the right hand of above equation to $\mathbf{K} + \varepsilon \mathbf{I}$ where ε is a small positive value. We added the differential penalty of the latent variables $\mathbf{t}$ to the constraint condition of kernel PCA. Kernel smoothed PCA is formulated as follows:

$$\begin{aligned} &\max \quad \mathrm{var}(\mathbf{t}) \\ &\text{s.t.} \quad \boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} + \kappa(\mathbf{Dt})'(\mathbf{Dt}) = 1 \end{aligned}$$

Finally, kernel smoothed PCA is written as the following generalized eigenvalue problem about the coefficients vector α:

$$\frac{1}{N}\mathbf{K}^2\boldsymbol{\alpha} = \lambda(\mathbf{K} + \kappa\mathbf{K}\mathbf{D}'\mathbf{D}\mathbf{K})\boldsymbol{\alpha}$$

α_i is coefficient vector corresponds to the large eigenvalue in i-th. We can calculate the latent variables $\mathbf{t}_i$ as $\mathbf{t}_i= \mathbf{K}\alpha_i$.

2.4.2. Kernel PLS

Rosipal et al. [13] introduced kernel PLS using a nonlinear iterative partial least squares (NIPALS) algorithm. As well as PLS, kernel PLS is explained as the optimization problem of maximizing the covariance between the latent variable **t**, which is a linear combination of the explanatory variable Φ in the feature space, and the latent variable **s**, which is a linear combination of the response variable **Y.**

$$\begin{aligned} &\max \quad \mathrm{cov}(\mathbf{t},\mathbf{s}) \\ &\text{s.t.} \quad \boldsymbol{\alpha}_x'\mathbf{K}\boldsymbol{\alpha}_x = 1 \quad \mathbf{w}_y'\mathbf{w}_y = 1 \end{aligned}$$

Finally, kernel PLS is written as the following generalized eigenvalue problem about the coefficients vector α_x:

$$\frac{1}{N^2}\mathbf{KYY'K}\boldsymbol{\alpha}_x = \lambda\mathbf{K}\boldsymbol{\alpha}_x$$

We added the differential penalty of the latent variables **t** to the constraint condition of kernel PLS. Kernel smoothed PLS is formulated as follows:

$$\begin{aligned} &\max \quad \mathrm{cov}(\mathbf{t},\mathbf{s}) \\ &\text{s.t.} \quad \boldsymbol{\alpha}_x'\mathbf{K}\boldsymbol{\alpha}_x + \kappa(\mathbf{Dt})'(\mathbf{Dt}) = 1 \quad \mathbf{w}_y'\mathbf{w}_y = 1 \end{aligned}$$

Finally, kernel smoothed PLS is written as the following generalized eigenvalue problem about the coefficients vector α_x:

$$\frac{1}{N^2}\mathbf{KYY'K}\boldsymbol{\alpha}_x = \lambda(\mathbf{K} + \kappa\mathbf{KD'DK})\boldsymbol{\alpha}_x$$

2.4.3. Kernel OPLS

As well as kernel PLS, kernel OPLS is explained as the optimization problem of maximizing the covariance between the latent variable **t**, which is a linear combination of the explanatory variable Φ in the feature space, and the latent variable **s**, which is a linear combination of the response variable **Y.** The constraint condition of kernel OPLS is different from kernel PLS.

$$\max \quad \mathrm{cov}(\mathbf{t},\mathbf{s})$$
$$\text{s.t.} \quad \boldsymbol{\alpha}_x'\mathbf{K}\boldsymbol{\alpha}_x = 1 \quad \mathbf{s}'\mathbf{s} = 1$$

Finally, kernel OPLS is written as the following generalized eigenvalue problem about the coefficients vector α_x:

$$\frac{1}{N^2}\mathbf{KY}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{K}\boldsymbol{\alpha}_x = \lambda\mathbf{K}\boldsymbol{\alpha}_x$$

We added the differential penalty of the latent variables **t** to the constraint condition of kernel smoothed OPLS. Kernel smoothed OPLS is formulated as follows:

$$\max \quad \mathrm{cov}(\mathbf{t},\mathbf{s})$$
$$\text{s.t.} \quad \boldsymbol{\alpha}_x'\mathbf{K}\boldsymbol{\alpha}_x + \kappa(\mathbf{Dt})'(\mathbf{Dt}) = 1 \quad \mathbf{s}'\mathbf{s} = 1$$

Finally, kernel smoothed OPLS is written as the following generalized eigenvalue problem about the coefficients vector α_x:

$$\frac{1}{N^2}\mathbf{KY}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{K}\boldsymbol{\alpha}_x = \lambda(\mathbf{K} + \kappa\mathbf{KD}'\mathbf{DK})\boldsymbol{\alpha}_x$$

2.4.4. Kernel FDA

Akaho [14] and Bach and Jordan [15] introduced kernel CCA, which is a kernelized version of RCCA. Kernel FDA [16] is introduced by using l_1-norm penalty to solve the sparse solution for the coefficient vector. Kernel FDA in this study is different from kernel FDA by Mika et al. In this section as well as the section 2.2.5, we explained kernel FDA by using the formulation of kernel CCA. Kernel FDA is formulated as follows:

$$\max \quad \mathrm{cov}(\mathbf{t},\mathbf{s})$$
$$\text{s.t.} \quad (1-\tau)\,\mathrm{var}(\mathbf{t}) + \tau\boldsymbol{\alpha}_x'\mathbf{K}\boldsymbol{\alpha}_x = 1 \quad \mathrm{var}(\mathbf{s}) = 1$$

Finally, kernel FDA is formulated as the following generalized eigenvalue problem about the coefficients vector α_x:

$$\frac{1}{N^2}\mathbf{KY}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{K}\boldsymbol{\alpha}_x = \lambda\left\{(1-\tau)\frac{1}{N}\mathbf{K}^2 + \tau\mathbf{K}\right\}\boldsymbol{\alpha}_x$$

We added the differential penalty of the latent variables **t** to the constraint condition of kernel FDA. Kernel smoothed FDA is formulated as follows:

$$\begin{aligned}&\max \quad \mathrm{cov}(\mathbf{t},\mathbf{s})\\ &\text{s.t.} \quad (1-\tau)\,\mathrm{var}(\mathbf{t}) + \tau\boldsymbol{\alpha}_x'\mathbf{K}\boldsymbol{\alpha}_x + \kappa(\mathbf{Dt})'(\mathbf{Dt}) = 1 \quad \mathbf{w}_y'\mathbf{w}_y = 1\end{aligned}$$

Finally, kernel smoothed FDA is written as the following generalized eigenvalue problem about the coefficients vector α_x:

$$\frac{1}{N^2}\mathbf{KY}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{K}\boldsymbol{\alpha}_x = \lambda\left\{(1-\tau)\frac{1}{N}\mathbf{K}^2 + \tau\mathbf{K} + \kappa\mathbf{KD}'\mathbf{DK}\right\}\boldsymbol{\alpha}_x$$

All these generalized eigenvalue problem can be computed easily by using Cholesky decomposition and singular value decomposition.

3. Data

3.1. Experimental data

S. cerevisiae MT8-1/plUX1X2XK strain developed in our previous study [17] was adapted to xylose. Three strains, the native strain, strain adapted to xylose under aerobic condition and anaerobic condition, were pre-cultured in 5 ml SD medium (20 g of glucose per liter, 6.7 g of yeast nitrogen base without amino acids [Difco Laboratories, Detroit, Mich.] per liter). After pre-culture, these strains were cultivated in 500ml SDC medium (SD medium with 20 g of casamino acids [Difco] per liter). The cells were washed and ethanol fermentation was carried out in 100 ml SX (50 g of xylose per liter, 6.7 g of yeast nitrogen base without amino acids per liter, 20 g of casamino acids per liter) medium. We collected the samples in triplicate at 0, 8, 16, 36, 60, 96 hours. The intracellular metabolites were analyzed by CE (Beckman Coulter, Fullerton, CA, USA)-MS (Applied Biosystems, Foster City, CA, USA) and GC (Agilent

Technologies, Wilmington, DE, USA)-TOFMS (Leco, St.Joseph, MI,USA).

Almost all intracellular metabolites were identified and the concentration of each metabolite is collected in data matrix with each element. The number of observations was 53 and the number of variables was 114. Each sample was categorized into three classes, native strain (class 1), the adaptated strain by xylose under aerobic condition (class 2), and under anaerobic condition (class 3). The number of samples with each class was 18, 18, and 17, because one sample was lost by the experimental error.

3.2. Software

The computer program in this study was developed in-house with the personal computer version of MATLAB 7 (Mathworks, Natick, MA, USA) by the authors. Our computer system to run these programs in this study has 2.4GHz CPU and 512MB size memory.

4. Results and discussion

4.1. Results of PCA, PLS, OPLS, RFDA with smoothness

A visualization result by using PCA, PLS, OPLS, and RFDA in which the high-dimensional data is projected onto 2-dimensional subspace was shown in Fig. 1. In Fig. 1, the symbol ○ is the native strain, the symbol ▽ is the adapted strain under aerobic condition and the symbol □ is the adapted strain under anaerobic condition. Supervised dimensionality reduction, PLS, OPLS, and RFDA, achieved separability between class well compared with PCA. And the higher separability was achieved by using RFDA than PLS and OPLS.

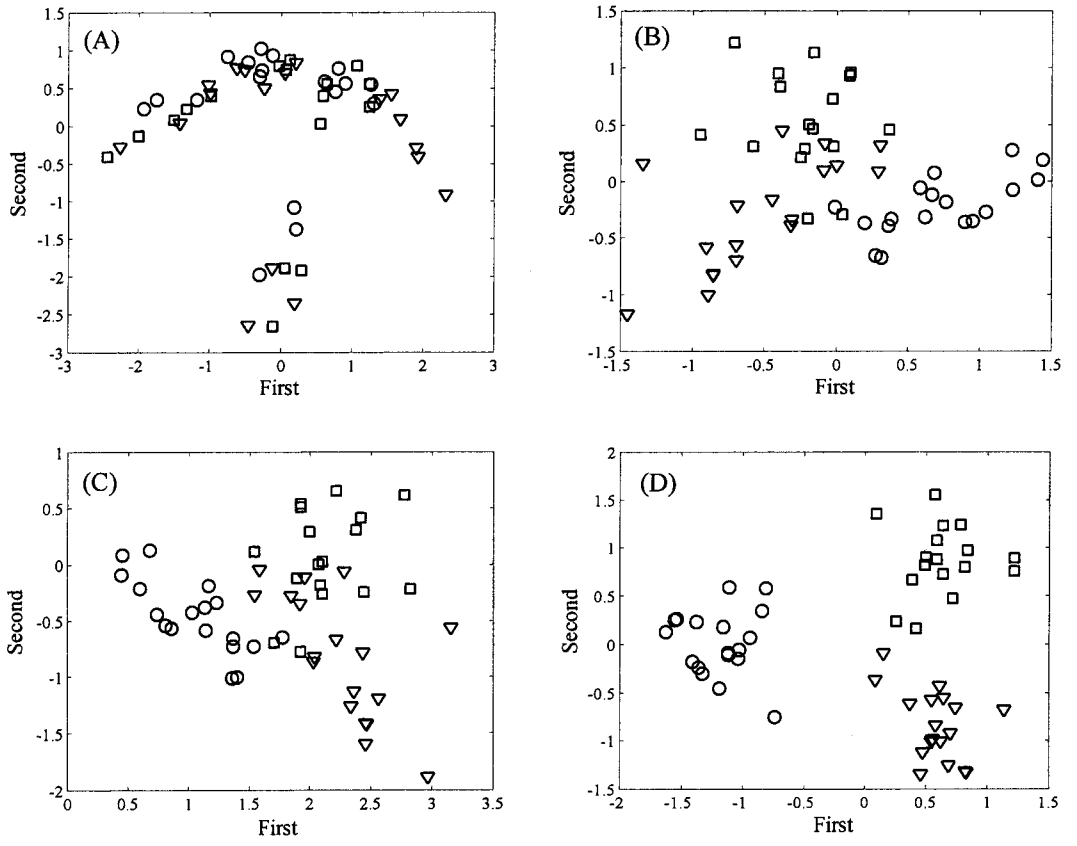

Fig. 1. Results of PCA, PLS, OPLS, and RFDA. The symbols (○, ▽, □) denote native strain (○), adapted strain under aerobic condition (▽), and under anaerobic condition (□). (A) Result of PCA. (B) Results of PLS. (C) Result of OPLS. (D) Result of RFDA ($\tau = 0.1$).

In this study, 1st and 2nd axis was used for visualization. Only two axes can be extracted in PLS, OPLS, and RFDA in the case of 3-class classification, whereas more than two axis can be extracted in PCA. The result of PCA is not always optimal for separability among classes because the purpose of PCA is to find good representation of original data in lower dimensional subspace. The result of PLS (Fig. 1 (B)) is similar to OPLS (Fig. 1 (C)). The matrix $(\mathbf{Y}'\mathbf{Y})^{-1}$ becomes identity matrix multiplied by constant, so we can eliminate the matrix $(\mathbf{Y}'\mathbf{Y})^{-1}$ from Eq. (3) when the number of samples within class is same. As a result, the Eq. (2) of PLS becomes equal to the Eq. (3) of OPLS. The regularized parameter of RFDA was set to τ=0.1. The methodology of setting the regularized parameter τ is described in the next section.

A perfect separability between the native strain (○) and the adapted strain (□, ▽) was achieved by RFDA in 1st axis. The 1st axis can be interpreted as the axis that is concerned with the adaptation. The effect of adaptation was known in ethanol fermentation from xylose [18]. In 2nd axis, the separability among the native strain (○) and the adapted strain under aerobic condition (□) and anaerobic condition (▽) was satisfactory. However, it is difficult to interpret the 2nd axis from biological point of view because the difference between adaptation under the aerobic condition (□) and anaerobic condition (▽) is not clear in experimental phase.

The results of smoothed PCA, PLS, OPLS, and RFDA were shown in Fig. 2. The smoothed parameter κ was set to 0.1. Compared to the results by original methods in Fig. 1, the effect of smoothness was found, especially in smoothed PCA. The 1st axis in smoothed PCA (Fig. 2 (A)) can be interpreted as the axis that is concerned with time change. The effect of differential penalty can be found in other methods but not in greater detail.

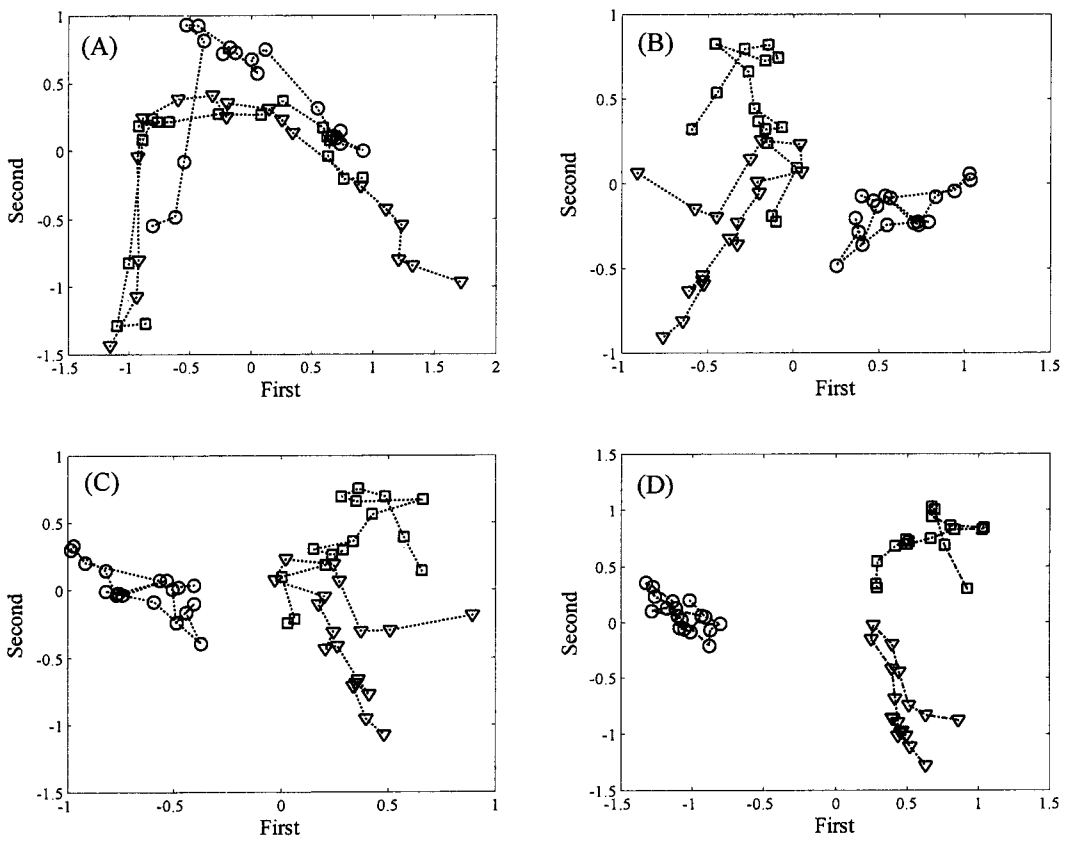

Fig. 2. Results of smoothed PCA, PLS, OPLS, and RFDA. The symbols are the same as Fig. 1. Each symbols within class was connected from 0 hour to 96 hour observations by the dotted line. (A) Result of smoothed PCA. (B) Results of smoothed PLS. (C) Result of smoothed OPLS. (D) Result of smoothed RFDA ($\tau = 0.1$).

4.2. Determination of regularized parameter of RFDA

The results of RFDA with different value of the regularized parameter (τ = 0.00001, 0.01, 0.1, 1) were shown in Fig. 3. As shown in Fig. 3, the degree of separation of classes increased with decrease in the value of regularized parameter τ, whereas separation of individual data points in each class was increased with increase in the regularized parameter τ. It was seen in Fig. 3 (A) and (B) that individual data points were strongly assembled when the regularized parameter τ is up to 0.01 and prediction is not possible because of the high dimensional data. Even though, there is a slight decrease in the separation of classes in Fig. 3 (C) and (D), the separation of individual data points allows the prediction of data. Therefore, it is necessary to optimize the regularized parameter τ. The regularized parameter τ in RFDA was determined by using leave-one-out cross-validation (LOOCV). We remove one sample as a test sample and compute the lower dimensional subspace by using the rest of the samples. In this subspace, which class a test sample belongs to was checked by using Mahalanobis distance. For all samples, this operation has been iterated. Sum of the number of misclassification against the regularized parameter τ was plotted in Fig. 4. The regularized parameter corresponding to the minimum value of the sum of the number of misclassification is optimal.

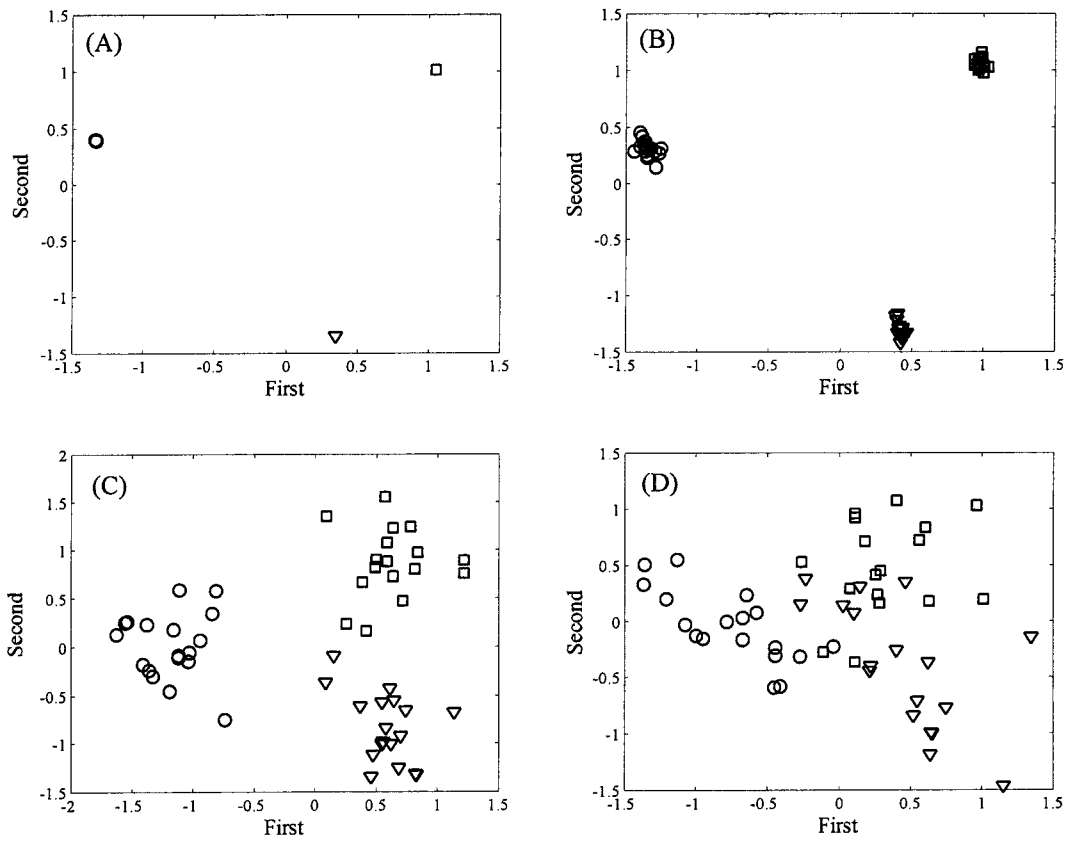

Fig. 3. Results of RFDA with different regularized parameter τ. The symbols are the same as Fig. 1. (A) Result of smoothed RFDA (τ = 0.00001). (B) Result of smoothed RFDA (τ = 0.01). (C) Result of smoothed RFDA (τ = 0.1). (D) Result of smoothed RFDA (τ = 1).

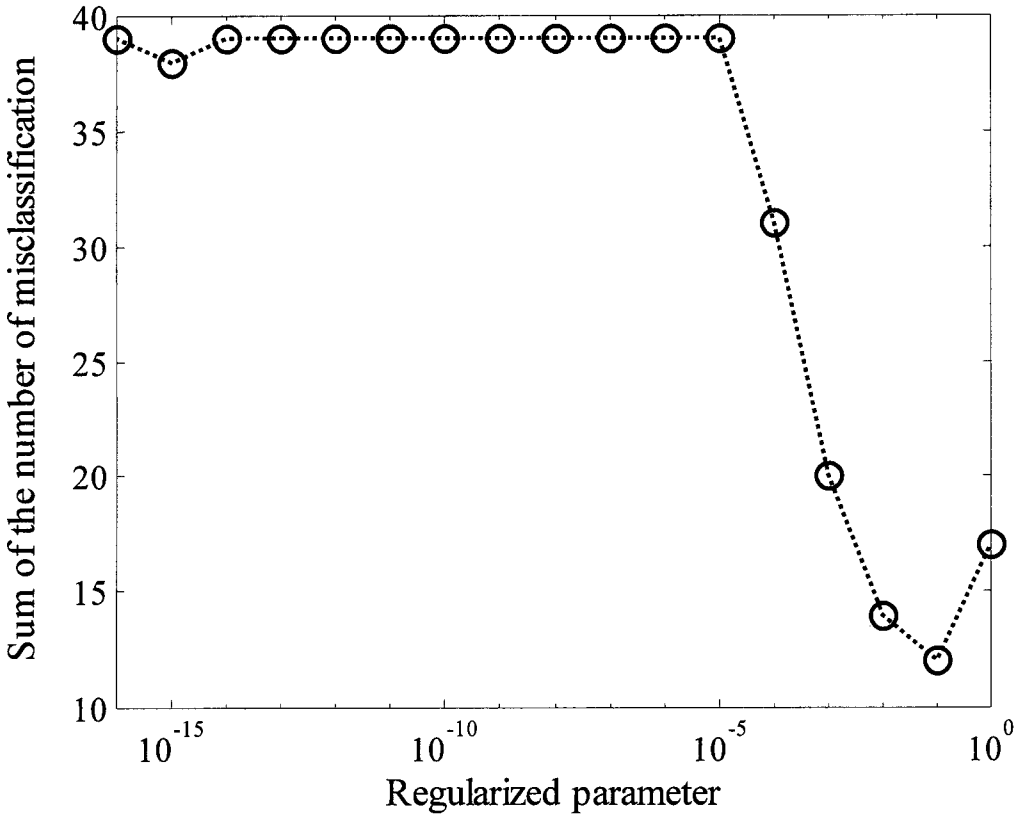

Fig. 4. Plot of leave-one-out cross-validation (LOOCV).

4.3. Kernel methods with smoothness

In this study, high separability among class is achieved by using linear dimensionality reduction. Therefore, the results of nonlinear dimensionality reduction were shown only as a guide. The results of kernel PCA, kernel PLS, kernel OPLS, and kernel FDA were shown in Fig. 5. As a kernel function, 2nd order polynomial kernel was applied. The smoothed parameter κ was set to 0.1. Separability among class is not achieved (Fig. 5 (A), (B), (C)) except kernel FDA (Fig. 5 (D)). However, the results of kernel FDA are not satisfactory for the same reason as RFDA in the section 4.2. The results of kernel smoothed PCA, PLS, OPLS, and FDA were shown in Fig. 6, and the effect of smoothness can be found. The high separability among class was achieved by using kernel smoothed PLS and OPLS compared with kernel PLS and OPLS.

When we apply RFDA (RCCA) to large-size data such as high-dimensional data, it often requires a large amount of memory and long calculation time. In such case, kernel FDA with linear kernel as a kernel function can reduce the calculation time drastically. In this study, a marked difference of calculation time could not be found because the data size is comparatively small.

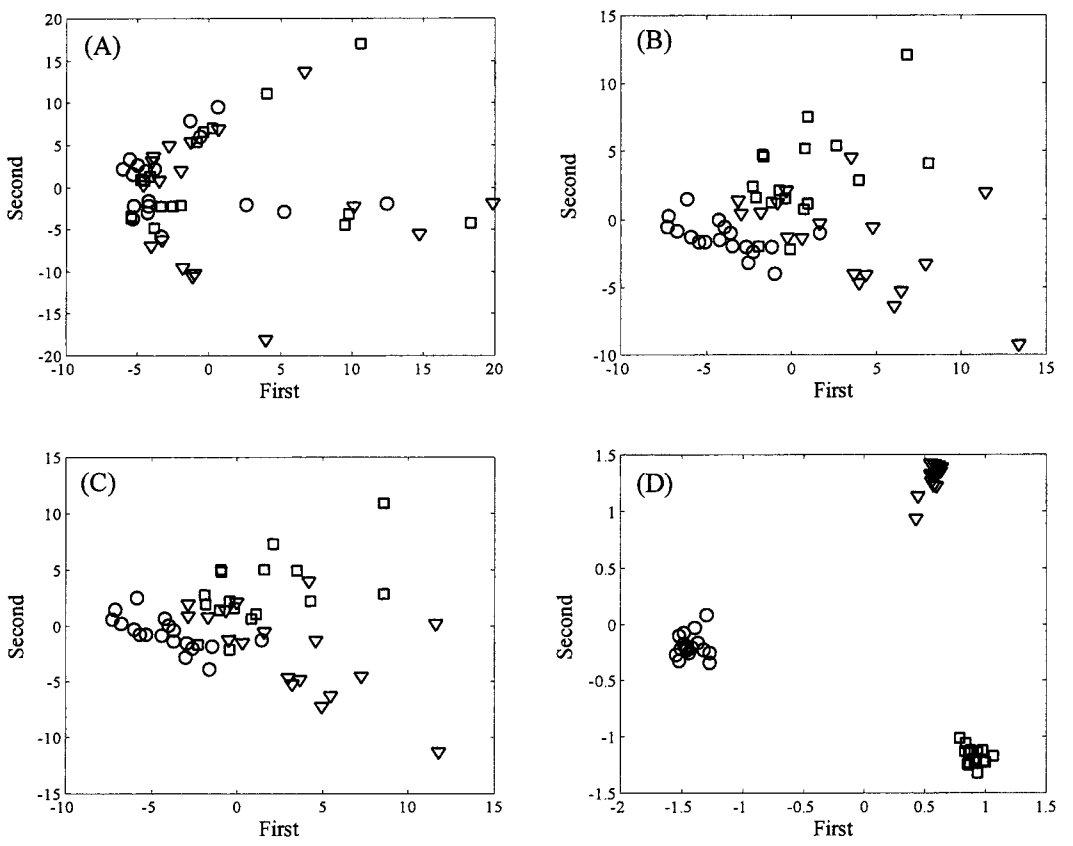

Fig. 5. Results of kernel PCA, PLS, OPLS, and FDA with polynomial kernel (d = 2). The symbols are the same as Fig. 1. (A) Result of kernel PCA. (B) Results of kernel PLS. (C) Result of kernel OPLS. (D) Result of kernel FDA ($\tau = 0.1$).

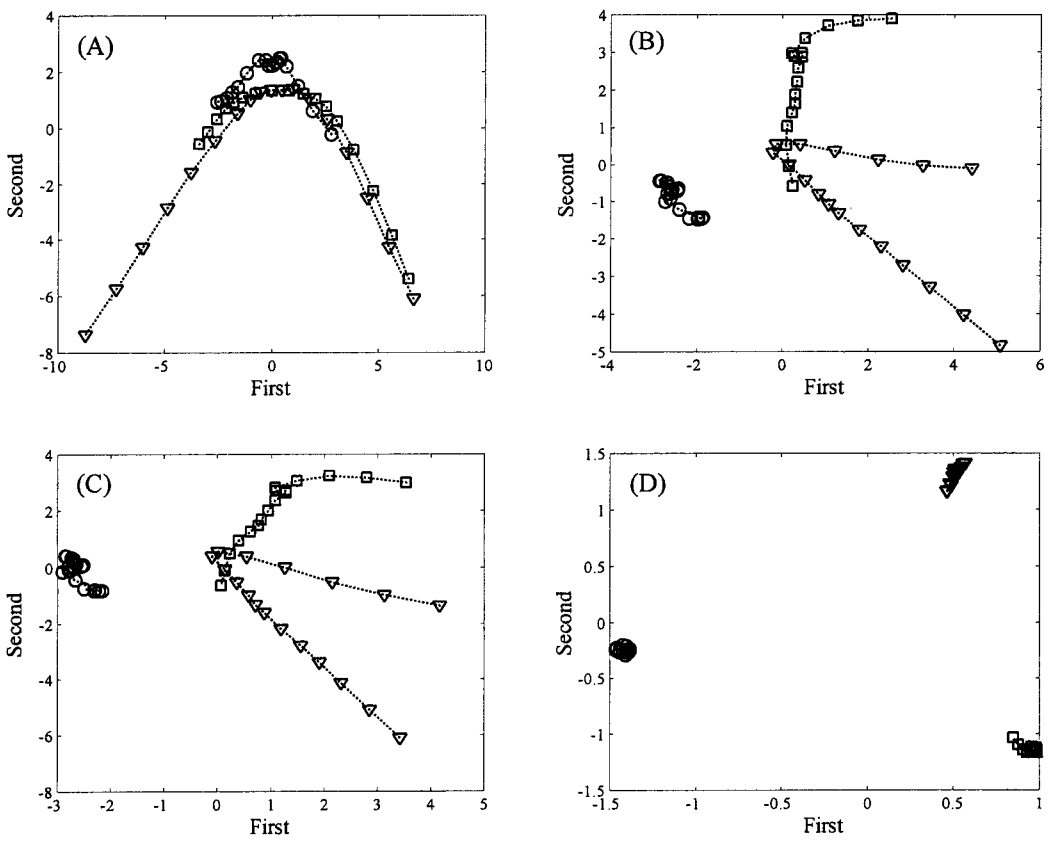

Fig. 6. Results of kernel smoothed PCA, PLS, OPLS, and FDA with polynomial kernel (d = 2). The symbols and lines are the same as Fig. 2. (A) Result of kernel smoothed PCA. (B) Results of kernel smoothed PLS. (C) Result of kernel smoothed OPLS. (D) Result of kernel smoothed FDA ($\tau = 0.1$).

References

[1] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom., 17 (2001) 166–173.

[2] K. Fukunaga, Introduction to Statistical Pattern Recognition (2ed.), Academic Press, Inc., Boston, 1990.

[3] B. W. Silverman, Smoothed functional principal components analysis by choice of norm, Ann. Statist., 24 (1996) 1–24.

[4] Z.P. Chen, J.H. Jiang, Y. Li, H.L. Shen, Y.Z. Liag, R.Q. Yu, Determination of the number of components in mixtures using a new approach incorporating chemical information , Anal. Chim. Acta, 13 (1999) 15–30.

[5] J.O. Ramsay, B.W. Silverman, Functional Data Analysis (1ed.), Springer, New York, 1997.

[6] R. Rosipal, N. Krämer, Overview and recent advances in partial least squares, in C. Saunders, M. Grobelnik, S. Gunn, J. Shawe-Taylor (Eds.), Subspace, Latent Structure and Feature Selection Techniques, Springer, New York, 2006, pp. 34–51.

[7] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, 2004.

[8] P.H.C. Eilers, A perfect smoother, Anal. Chem., 75 (2003) 3631–3636.

[9] N. Krämer, A.L. Boulesteix, G. Tutz, Penalized partial least squares based on b-splines transformations, SFB 386, Discussion Paper 483 (2007)

[10] U. Indahl, H. Martens, T.J. Næs, From dummy regression to prior probabilities in PLS-DA, J. Chemom., 21 (2007) 529–536.

[11] H. Yamamoto, H. Yamaji, E. Fukusaki, H. Ohno, H. Fukuda, Canonical correlation analysis for multivariate regression and its application to metabolic fingerprinting, Biochem. Eng. J. (in press).

[12] B. Schölkopf, A.J. Smola, K.R. Müller, Nonlinear component analysis as a kernel

eigenvalue problem, Neural. Comp., 10 (1998) 1299–1319.

[13] R. Rosipal, L.J. Trejo, Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space, JMLR, 2 (2001) 97–123.

[14] S. Akaho, A kernel method for canonical correlation analysis, International Meeting of Psychometric Society, 2001.

[15] F.R. Bach, M.I. Jordan, Kernel independent component analysis, JMLR, 3 (2002) 1–48.

[16] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K.R. Müller, Fisher discriminant analysis with kernels, in Y.-H. Hu, J. Larsen, E. Wilson and S. Douglas (Eds.), Neural Networks for Signal Processing IX, August, 1999, Madison, USA, IEEE, New York, 1999, pp. 41–48.

[17] S. Katahira, Y. Fujita, A. Mizuike, H. Fukuda, A. Kondo, Construction of a Xylan-Fermenting Yeast Strain through Codisplay of Xylanolytic Enzymes on the Surface of Xylose-Utilizing *Saccharomyces cerevisiae* Cells, Appl. Environ. Microbiol., 70, (2004) 5407–5014.

[18] M. Kuyper, M.J. Toirkens, J.A. Diderich, A.A. Winkler, J.P. van Dijken, J.T. Pronk, Evolutionary engineering of mixed-sugar utilization by a xylose-fermenting Saccharomyces cerevisiae strain, FEMS Yeast Res., 5, (2005) 925–934.

Part III

Multivariate analysis for curve resolution

Application of regularized alternating least squares and independent component analysis to curve resolution problem

1. Introduction

Metabolomics represents the exhaustive profiling of the metabolites contained in organisms. The final goal of metabolomics is the profiling of all metabolites contained in the target organism. However, it is almost impossible to perform profiling of as many as ten thousand metabolites due to lack of appropriate technology, and further development is required at each of the steps involved in metabolomics analysis, which consists of cultivation, sampling, extraction, data matrix conversion, data mining, and bioscience feedback [1].

To detect a number of metabolites, various analytical equipment has been used. In the present study, we considered only hyphenated techniques such as GC-MS and LC-diode array detection (DAD). In chromatography, complete separation of all metabolites is desirable to maintain a high degree of quantification, but several important metabolites may be coeluted because samples derived from living organisms contain many complex metabolites.

Self-modeling curve resolution (SMCR) [2] is a chemometrics technique by which separation is achieved. SMCR has recently been applied extensively to two-way data obtained from chromatography hyphenated with multi-channel detection methods, for instance GC-MS and HPLC-DAD. In principle, SMCR does not require *a priori* information about the pure components. The only pre-requirement is generic knowledge about the pure variables such as non-negativity or unimodality.

Alternating least squares (ALS) [3] is a classical algorithm of SMCR and has recently been applied in metabolomics studies [4]. ALS starts iteration with randomly chosen concentration vectors for the pure variables. By its character, ALS produces different solutions depending on the initial values, which presents a problem in practical

use. This problem is often avoided by applying multivariate curve resolution to ALS, which uses initial estimate that depends on the analysis equipment [5]. In the present study, we sought to avoid the problem by using a normalized constraint in calculating ALS. We also improved the ALS algorithm by adding a regularized term.

Independent component analysis (ICA) [6] is comparatively new and mainly studied in branches of information science, such as signal processing and neural networks. Like SMCR, ICA does not require *a priori* information about the pure components, but only statistical independence among them. In chemometrics studies, the application of ICA to near-infrared spectroscopy has been studied [7], but its application to curve resolution problems is still very infrequent.

In the present study, we applied two curve resolution methods, the regularized version of ALS (RALS) and ICA, to the metabolite data of the photosynthetic microalga *Haematococcus pluvialis*, which contain a number of metabolites. We focus on astaxanthin, a substance with high antioxidant action which is produced by *H. pluvialis*, accumulated at high concentration in the cells, and coeluted in chromatography. In an assessment of the effectiveness of these curve resolution methods, we show that they are useful in the case of coelution in chromatography, as with metabolome data, and for identification of metabolites.

2. Experimental

2.1. Microorganism and cultivation conditions

H. pluvialis NIES-144, which was obtained from the Microbial Culture Collection in the National Institute for Environmental Studies (Tsukuba, Japan), was grown aerobically under illumination by light emitting diodes (Toyoda Gosei, Aichi, Japan) in Kobayashi's basal medium (pH 6.8) containing 0.12 w/v% sodium acetate, 0.2 w/v% yeast extract, 0.04w/v% L-asparagine, 0.002 w/v% $MgCl_2 \cdot 6H_2O$, 0.001 w/v%

$FeSO_4 \cdot 7H_2O$, and 0.002 w/v% $CaCl_2 \cdot 2H_2O$ [8]. After precultivation in a 200-cm^3 Erlenmeyer flask set in a glass-sided water bath under illumination at 3.8 μmol-photon m^{-2} s^{-1} with a fluorescent lamp, *H. pluvialis* was inoculated at 0.03 mg-dry cell cm^{-3} into a 50-cm^3 Erlenmeyer flask containing 50 cm^3 of the medium, or into a glass vessel of 6.5 cm height, 5.0 cm width and 2.6 cm depth with a working volume of 55 cm^3, and cultivated at 20°C under agitation with a magnetic stirrer bar.

2.2. Measurement of metabolites with focus on astaxanthin in cells

Aliquots of 0.5 cm^3 of the cell suspension were removed from the culture vessel and centrifuged at 7500 × g for 10 min. The cell precipitate was resuspended in 0.5 cm^3 of methanol and mixed with 0.40 g of silica particles (particle diameter, 0.2–1 mm; Kanto Chemical, Tokyo, Japan). To extract astaxanthin and metabolites from the cells, the mixture was treated vigorously with a vortex mixer for 10 min and centrifuged for 10 min. The supernatant (0.5 cm^3) was mixed with 0.1 cm^3 of a methanol solution of NaOH (5 mM NaOH) and kept at 20°C overnight under nitrogen and in darkness. The metabolites in the prepared samples were measured with an HPLC system (LC-10; Shimadzu, Kyoto, Japan) equipped with a reverse phase column (Cosmosil 5C18-MS-II, 4.6×150 mm; Nacalai Tesque, Kyoto, Japan). The mobile phase was methanol with a flow rate of 1 cm^3 min^{-1} and the absorbance of the effluent solution was measured from 350 nm to 700 nm with a photodiode array detector (SPD-M10A; Shimadzu).

3. Data analysis

3.1. Synthesized dataset

We synthesized an artificial dataset, which was generated by means of $\mathbf{X} = \mathbf{CS}$ where $\mathbf{X}$ was the data matrix, $\mathbf{S}$ was the two Gaussian type functions shown in Fig. 1A,

Table 1. The initial value of **C** in artificial dataset

	Component 1	Component2
Sample 1	0.0956	0.8961
Sample 2	0.0694	0.6539
Sample 3	0.8151	0.8738
Sample 4	0.8808	0.1014
Sample 5	0.1588	0.7608

and **C** was the random value distributed on the interval [0, 1] given in Table 1.

3.2. HPLC-DAD dataset

A data matrix was formed from HPLC-DAD as follows:

$$\mathbf{X} = \begin{matrix} \text{Retention time (min)} \downarrow \\ \\ \end{matrix} \begin{matrix} \text{Wavelength (nm)} \rightarrow \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \end{matrix}$$

In this matrix, each row corresponds to the spectrum at a certain retention time and each column to the chromatogram at a certain wavelength. **C** and **S**, respectively, denote the matrix of the concentration profile and of the pure spectrum of each metabolite. The following bilinear equation is assumed:

$$\mathbf{X} = \mathbf{CS}$$

SMCR decomposes the data matrix **X** to **C** and **S** without *a priori* information. In HPLC-DAD datasets, each column of **C** corresponds to the chromatogram of the pure components and each row of **S** to the diode array spectrum of the pure components.

3.3. Data analysis methods

3.3.1. Determination of component number

Before we apply the curve resolution methods, we must estimate the number of pure components. In the present paper, we used the ratio of singular values (SVR) [9]. SVR extracts information about selectivity in singular value evolving profiles in order to more effectively determine peak homogeneity.

3.3.2. Alternating least squares (ALS)

ALS is a classical SMCR algorithm for decomposing the data matrix **X** into the two matrices **S** and **C** under non-negative constraints, as follows:

$$\mathbf{X} = \mathbf{CS} \qquad (1)$$

The algorithm comprises an iterative solving of two least squares problems. First, **S** is estimated from the original data matrix, **X**, and an assumed value of **C**, and then **C** is estimated from **X** and the estimated **S**, as follows:

$$\mathbf{S} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{X} \qquad (2)$$

$$\mathbf{C} = \mathbf{X}\mathbf{S}^T(\mathbf{S}\mathbf{S}^T)^{-1} \qquad (3)$$

These two-step linear calculations are carried out iteratively until an appropriate convergence condition, such as $||\mathbf{X} - \mathbf{CS}|| < \varepsilon$, is satisfied or the calculations are iterated 1000 times.

ALS algorithm is as follows:

1. **C** is initialized by using a normally distributed random value.
2. *S* is determined by using the data matrix **X** and **C**, as $\mathbf{S} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{X}$.
3. Negative elements of **S** are set to zero.
4. *C* is estimated using the data matrix **X** and **S**, as $\mathbf{C} = \mathbf{X}\mathbf{S}^T(\mathbf{S}\mathbf{S}^T)^{-1}$.
5. Negative elements of **S** are set to zero.

6. Repeat from Step 2 to 5 until $||\mathbf{X} - \mathbf{CS}||$ converges to ε or the calculations are repeated 1000 times.

3.3.3. Regularized ALS (RALS)

In the context of regression problems, the ridge regression method is suggested, which involves a regularized term [10]. We proposed a regularized version of ALS, RALS. This is implemented very easily by adding a regularizing parameter, λ. Thus, Eqs. (2) and (3) in ordinary ALS become the following modified equations

$$\mathbf{S} = (\mathbf{C}^T\mathbf{C} + \lambda\mathbf{I})^{-1}\mathbf{C}^T\mathbf{X} \qquad (4)$$

$$\mathbf{C} = \mathbf{X}\mathbf{S}^T(\mathbf{S}\mathbf{S}^T + \lambda\mathbf{I})^{-1} \qquad (5)$$

where **I** is the identity matrix.

Ordinary ALS is a special case of RALS in which the regularized parameter λ is set to zero.

3.3.4. Normalized constraint

We used a constraint which normalizes the vector norm of each row of **S** to 1. This constraint was implemented in the (R)ALS algorithm between Step 2 and Step 3.

3.3.5. Independent component analysis (ICA)

ICA decomposes the data matrix **X** into **C** and **S**, assuming only that **C** and **S** are statistically independent. The statistical independence can be reflected by non-Gaussianity. In fast ICA algorithms [11], non-Gaussianity is measured by kurtosis, in its generalization, negentropy. The ICA algorithm is designed so that the number of samples is equal to the number of independent components. Otherwise, principal component analysis (PCA) is widely used for the dimension reduction of a data matrix.

Suppose that we seek for one of the independent components as $\mathbf{y}=\mathbf{w}^{\mathrm{T}}\mathbf{x}$ by using negentropy, where $\mathbf{y}$ is an independent component and $\mathbf{w}$ is a weight vector. Negentropy is based on the quantity of differential entropy. The larger the value of negentropy becomes, the more the variables are statistically independent. The differential entropy H of a random variable y with density $p(\mathbf{y})$ is defined as:

$$H(\mathbf{y}) = -\int p(\mathbf{y}) \log p(\mathbf{y}) d\mathbf{y}$$

The negentropy J of $\mathbf{y}$ is then defined as:

$$J(\mathbf{y}) = H(\mathbf{y}_{\mathbf{gauss}}) - H(\mathbf{y})$$

where $\mathbf{y}_{\mathbf{gauss}}$ is a Gaussian random vector of the same covariance matrix as $\mathbf{y}$. The estimation of independent components is reduced to finding $\mathbf{y}$ in a way that maximizes $J(\mathbf{y})$.

The approximation of negentropy takes the form:

$$J(\mathbf{y}) \approx [E\{G(\mathbf{y})\} - E\{G(\nu)\}]^2$$

where ν is the normalized Gaussian variable, $\mathbf{y}$ is assumed to have zero mean and unit variance, G is a non-quadratic function, and E is expectation operator. Practical choices of G have been proposed as follows:

$$G_1(u) = \frac{1}{a_1} \log \cosh(a_1 u)$$

$$G_2(u) = -\frac{1}{a_2} \exp(-a_2 u^2 / 2)$$

$$G_3(u) = \frac{1}{4} u^4$$

where a_1 and a_2 are suitable positive constants. The maximum value of negentropy J_G

$$J_G(\mathbf{w}) = [E\{G(\mathbf{w}^{\mathrm{T}}\mathbf{x})\} - E\{G(\nu)\}]^2$$

under constraints of $E[(\mathbf{w}^{\mathrm{T}}\mathbf{x})^2] = 1$ would be likely to be found. We can solve this problem with Lagrangean or Newton-like iteration methods using the derivative function g of G. Finally, we obtain the following update rule [9] as follows:

$$\mathbf{w} \leftarrow E\{\mathbf{x}g(\mathbf{w}^{\mathrm{T}}\mathbf{x})\} - E\{g'(\mathbf{w}^{\mathrm{T}}\mathbf{x})\}\mathbf{w}$$

$$\mathbf{w} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

3.4. Software

The computer program for RALS calculation was developed in-house with the personal computer version of MATLAB 5.3 (Mathworks, Natick, MA, USA) by the authors. These codes can be run on a compatible PC. For ICA calculation, we used the FastICA package for MATLAB developed at the Helsinki University of Technology.

4. Results and discussion

4.1.Comparison of solution between ALS and ALS with normalized constraint for synthetic dataset

For the synthesized dataset shown in Fig. 1B, we applied ALS and ALS with a normalized constraint to the 100 different initial values of **C**. We overwrote the results of 100 simulations on a graph (Fig. 2). The figure shows that the number of solutions is greatly reduced by use of the normalized constraint.

The reason why the normalized constraint gives a reduced number of solutions is explained in the following way. In the matrix decomposition shown in Eq. (1), any diagonal matrix **M** can be inserted as follows:

$$\mathbf{X} = \mathbf{C}\mathbf{M}^{-1}\mathbf{M}\mathbf{S} = \tilde{\mathbf{C}}\tilde{\mathbf{S}}$$

This means that the matrix decomposition is not undertaken uniquely. The use of the normalized constraint is equivalent to setting any diagonal matrix **M** to

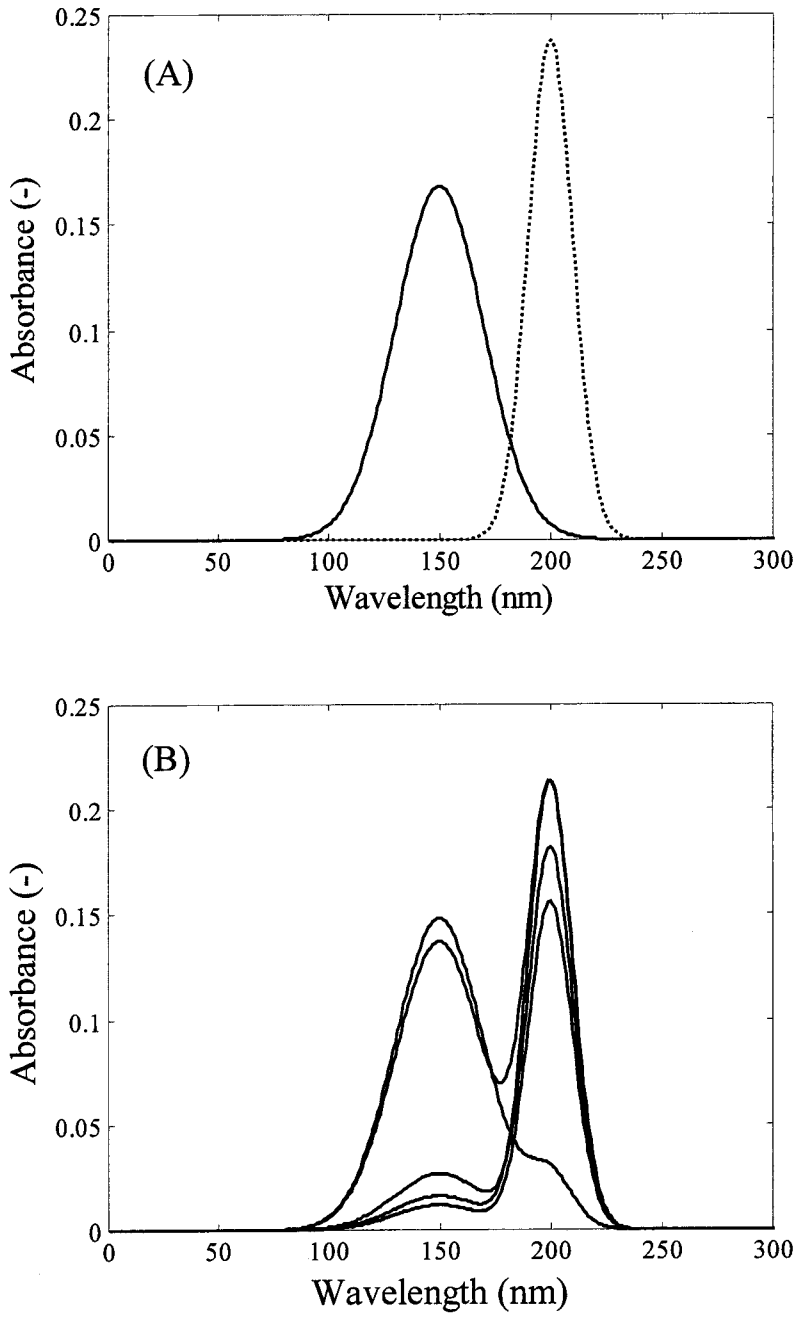

Fig. 1. Synthetic datasets. (A) Two Gaussian type functions as source spectra. The solid line is component 1 and the doted line component 2. (B) Synthetic data from source spectra mixed at a fixed ratio.

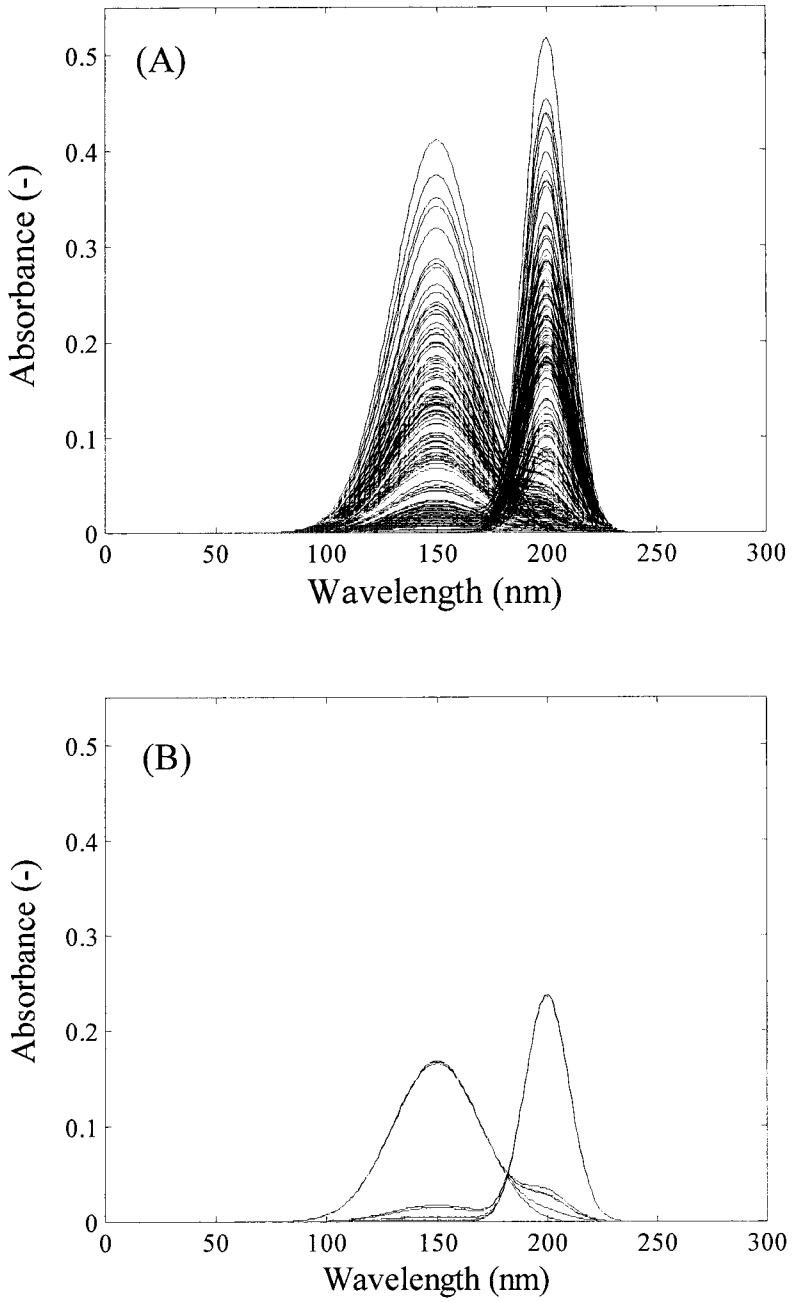

Fig. 2. Simulation results for alternating least squares (ALS) and ALS with normalized constraints using artificial dataset. (A) Results for ALS of 100 simulations with varying initial values. (B) Results for ALS with normalized constraints of 100 simulations with varying initial values.

$$M = \begin{bmatrix} 1/l_1 & 0 \\ 0 & 1/l_2 \end{bmatrix}$$

where l_i is the norm of the i-th row of **S**. The best choice of diagonal matrix **M** is a constraint physically meaningful for **C** or **S**. Given the constraint on either **C** or **S** regarding the norm, the solution of ALS is determined almost uniquely in such a simple case.

4.2. Robustness of RALS against inappropriate number of pure components

To show the robustness of the regularized version of ALS, known as RALS, we applied ALS and RALS to the 100 different initial values of **C** for the synthesized dataset shown in Fig. 1B. This dataset was artificially synthesized to contain two components. We calculated ALS and RALS on the assumption that the number of pure components is three, which is a dummy number. We overwrote the results of 100 simulations on a graph (Fig. 3). The figure shows that the number of solutions is greatly reduced by RALS, indicating that the robustness of RALS would be improved by adding a regularized term.

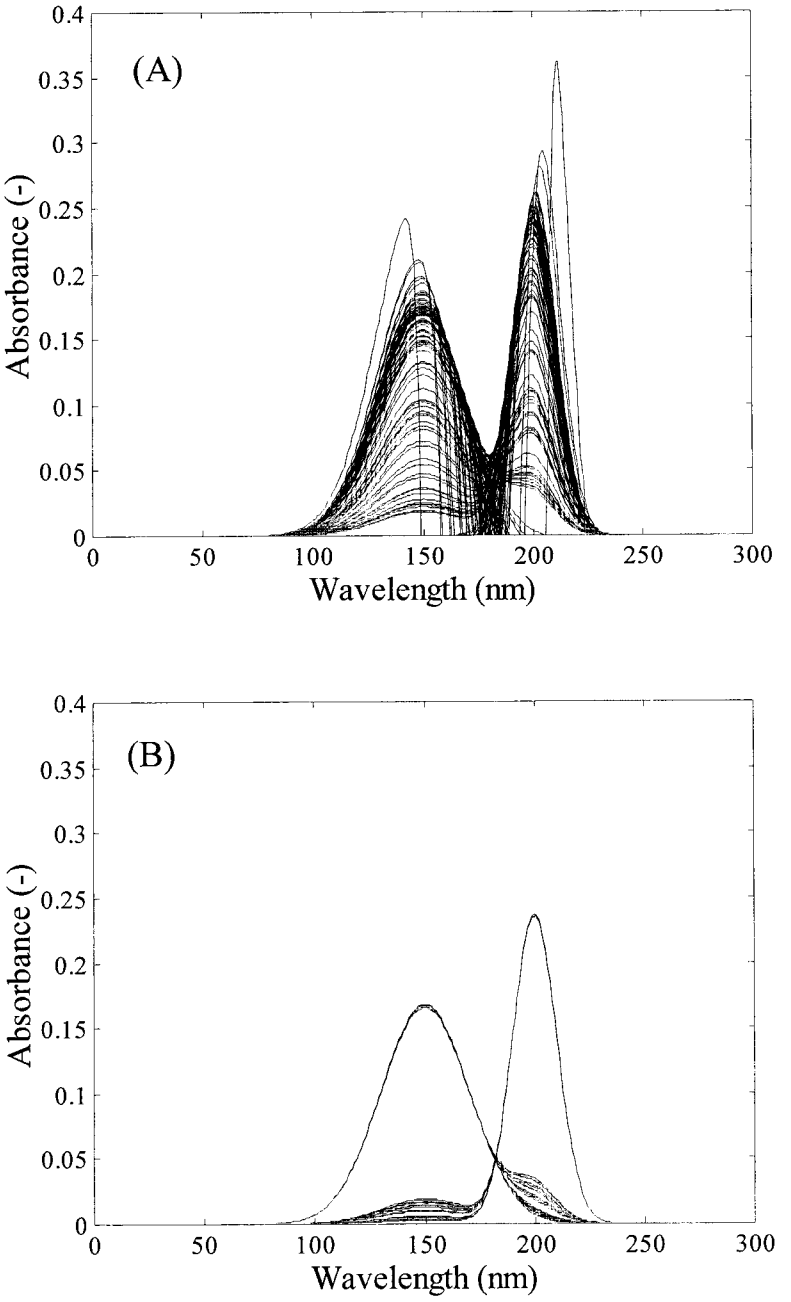

Fig. 3. Simulation results for ALS and regularized ALS (RALS) using artificial dataset. (A) Results for ALS of 100 simulations with varying initial values. (B) Results for RALS of 100 simulations with varying initial values. The value of regularized parameter λ is set to 0.0001.

4.3. Analysis of HPLC-DAD data of H. pluvialis metabolites

Before curve resolution for the real dataset, the HPLC-DAD data of *H. pluvialis* metabolites, we estimated the number of pure components and divided the whole dataset into subsets using SVR. Dividing the whole dataset into small subsets allows a complex problem to be made into a simple problem. Fig. 4 shows the results of SVR, which suggest that three subsets, data 1, data 2, and data 3, exist and that each subset involves only two pure components. In Fig. 5, the chromatogram and spectra of each subset are presented.

The results of curve resolution by (R)ALS with a normalized constraint and ICA are presented in Figs. 6 and 7, respectively. In the calculation of RALS, the value of the regularized parameter λ in the three datasets is 0, 0, and 0.01, respectively, as mentioned below. In ICA, each column of **C** in Eq. (1), which corresponds to the chromatogram, is considered to be independent of each other. We therefore seek the independent components of **C**, that is $\mathbf{C} = \mathbf{WX}$. The reason for this will be shown later. To make the comparison of ICA with RALS easy, the vector norm of each row of **S** was normalized to 1 in ICA. For data 1, inspection of the raw data did not show that the spectrum contained the spectra of two pure components. The curve resolution reveals that the data decomposed into two pure spectra and two chromatograms. As for data 1, data 2 and data 3 decompose into two pure spectra and two chromatograms. For data 2 and data 3, similar results of curve resolution were obtained by RALS and ICA. In the calculation of ICA, each independent component has unit variance. By its character, information on the concentration profiles in chromatogram has been lost. In Fig. 7 (D), negative absorbance was obtained for data 1, because ICA does not take the non-negativity constraint into consideration. For these reasons, the resolution accuracy becomes worse especially for data 1.

We used a normalized constraint in calculating RALS for the real dataset. As discussed above, the non-negativity constraint alone is not enough for practical use.

Because DAD shows the information for each pure component while information on the concentration profiles is consolidated in HPLC, the normalized constraint is valid for analyzing the HPLC-DAD dataset and is very comprehensible intuitively.

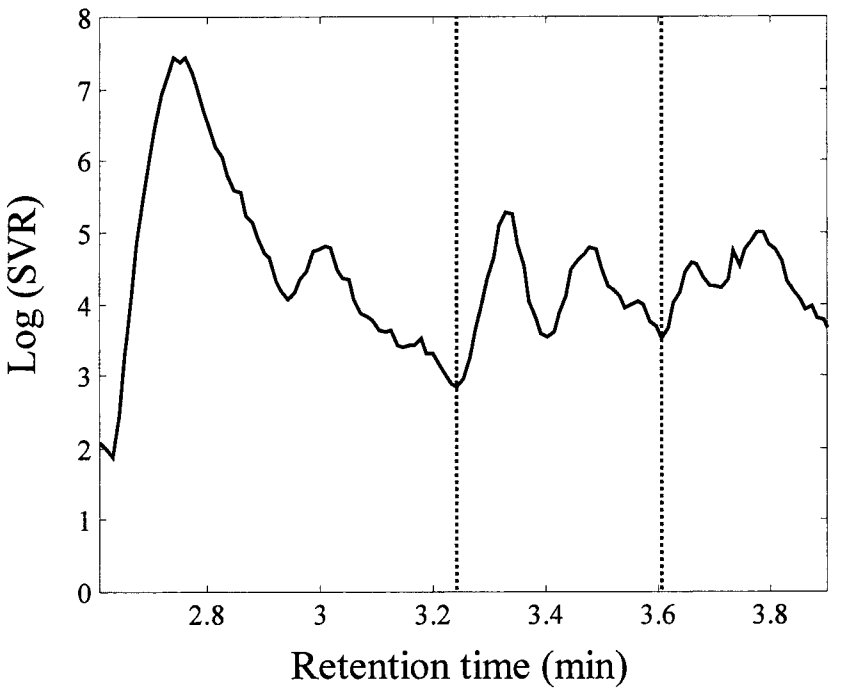

Fig. 4. Result of SVR. The vertical lines divide the whole dataset from 2.6 min to 3.9 min into three sub-datasets, data 1 from 2.6 min to 3.2 min, data 2 from 3.2 min to 3.6 min, and data 3 from 3.6 min to 3.9 min.

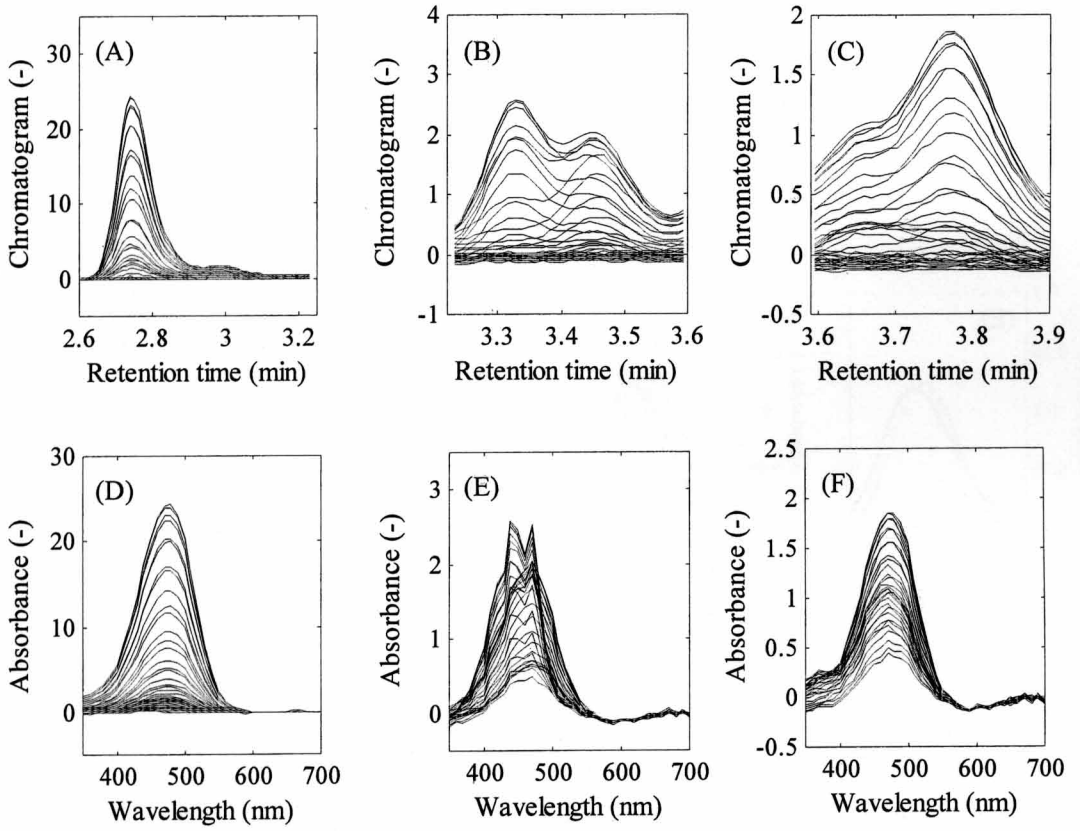

Fig. 5. Datasets for chromatogram and DAD spectra. (A) Chromatogram of data 1. (B) Chromatogram of data 2. (C) Chromatogram of data 3. (D) DAD spectrum of data 1. (E) DAD spectrum of data 2. (F) DAD spectrum of data 3.

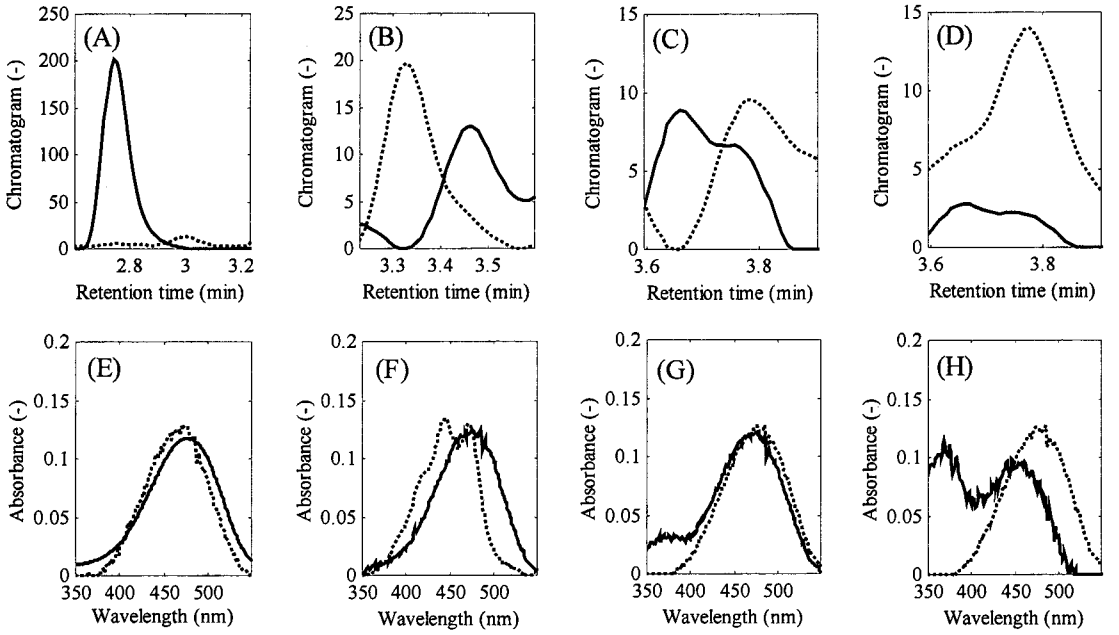

Fig. 6. Results of (R)ALS with normalized constraint. The solid line is one component and the doted line the other component. (A) Resolved chromatogram of data 1 by ALS. (B) Resolved chromatogram of data 2 by ALS. (C) Resolved chromatogram of data 3 by ALS. (D) Resolved chromatogram of data 3 by RALS. (E) Resolved DAD spectrum of data 1 by ALS. (F) Resolved DAD spectrum of data 2 by ALS. (G) Resolved DAD spectrum of data 3 by ALS. (H) Resolved DAD spectrum of data 3 by RALS. In RALS the value of regularized parameter λ is set to 0.01.

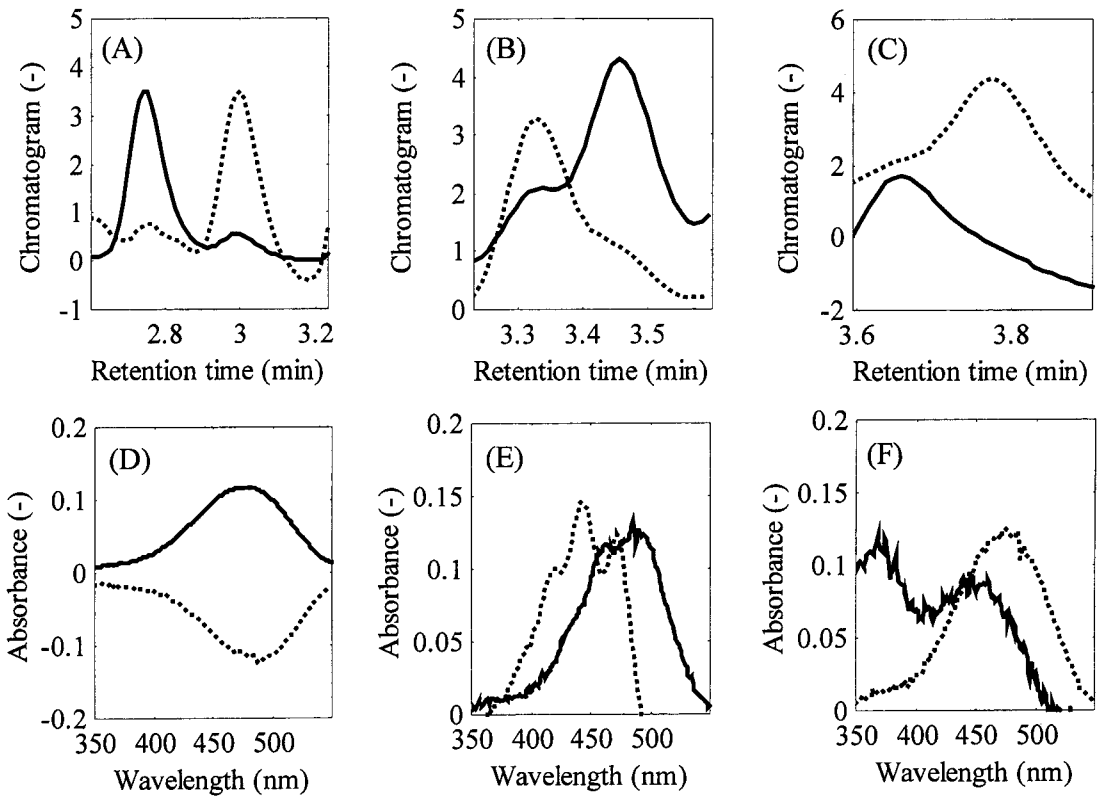

Fig. 7. Results of independent component analysis (ICA). The vector norm of each row of **S** was normalized to 1. The lines are the same as in Fig. 6. (A) Resolved chromatogram of data 1. (B) Resolved chromatogram of data 2. (C) Resolved chromatogram of data 3. (D) Resolved DAD spectrum of data 1. (E) Resolved DAD spectrum of data 2. (F) Resolved DAD spectrum of data 3

4.4. Identification of H. pluvialis metabolites

Table 2 presents a comparison of the position of the spectrum peaks obtained by each curve resolution method with those experimentally obtained by Yuan and Chen [12]. This table suggests that (3S,3'S)-*trans*-astaxanthin and adonirubin are contained in data 1, lutein and (3S,3'S)-9-*cis*-astaxanthin in data 2, and (3S,3'S)-13-*cis*-astaxanthin and (3R,3'R)-*trans*-astaxanthin in data 3. Table 2 shows that RALS and ICA successfully detected each metabolite peak in the spectra for data 3, whereas ALS failed in peak detection. It also suggests that regarding peak detection of each metabolite the results of RALS may be more accurate than those of ALS by using the optimal value of the regularized parameter λ.

In the detection of each metabolite peak in the spectra, high accuracy was achieved in comparison with the experimental results. Where our experimental conditions are different, it is of course difficult to compare directly with other experimental results or databases. The more sophisticated metabolomics becomes, the more mathematical methods will become necessary.

Table 2 Comparison of results of ALS, RALS and ICA with experimental results

	Metabolite	Peak wavelength (nm)			
		Experimental	ALS	RALS	ICA
Data 1	(3S,3'S)-*trans*-astaxanthin	480	477	477	476
	Adonirubin*	472.8	469	469	490
Data 2	Lutein	443.9, 472.8	445, 470	445, 470	442, 468
	(3S,3'S)-9-*cis*-astaxanthin	472.8	472	472	488
Data 3	(3S,3'S)-13-*cis*-astaxanthin	371.8, 472.8	474	369, 454	367, 447
	(3R,3'R)-*trans*-astaxanthin*	480	474	478	473

* Tentatively identified.

4.5. Determination of regularized parameter

In the present study, the regularized parameter λ was determined by using the Frobenius norm of $\mathbf{X} - \mathbf{CS}$, $||\mathbf{X} - \mathbf{CS}||$. For the three datasets, $||\mathbf{X} - \mathbf{CS}||$ was calculated by changing the value of λ from 10^{-10} to 1(Fig. 8). For data 1, the value of $||\mathbf{X} - \mathbf{CS}||$ increased almost exponentially. For data 2, $||\mathbf{X} - \mathbf{CS}||$ did not change at $\lambda < 0.01$ and then increased sharply. Hence we determined the value of the regularized parameter to be 0 for these two datasets. For data 3, we determined the optimal value of λ to be 0.01 because $||\mathbf{X} - \mathbf{CS}||$ was the minimum at $\lambda = 0.01$.

Here s_i denotes the vector of the i-th row of matrix **S**. If $E[s_i] = 0$ and $E[s_i^2] = 1$, the matrix $n^{-1}\mathbf{SS'}$, where n is the number of samples, is the correlation coefficient matrix. Hence the off-diagonal elements of the matrix **SS'** are deeply related to the correlation of the s_i and s_j which contains the information of the DAD spectra of pure components. Based on this consideration, we found that the resolution accuracy depends on the value of regularized parameter λ. When the regularized term and parameter are selected to be suitable for *a priori* information, the accuracy of the curve resolution will be better than when using ordinary ALS.

We improved the ALS algorithm by adding a regularized term. As can be seen in Fig. 6 (G) ($\lambda = 0$) and (H) ($\lambda = 0.01$), RALS was able to produce different solutions depending on the regularized parameter λ. It is therefore necessary to determine the optimal value of λ by investigating it for individual dataset.

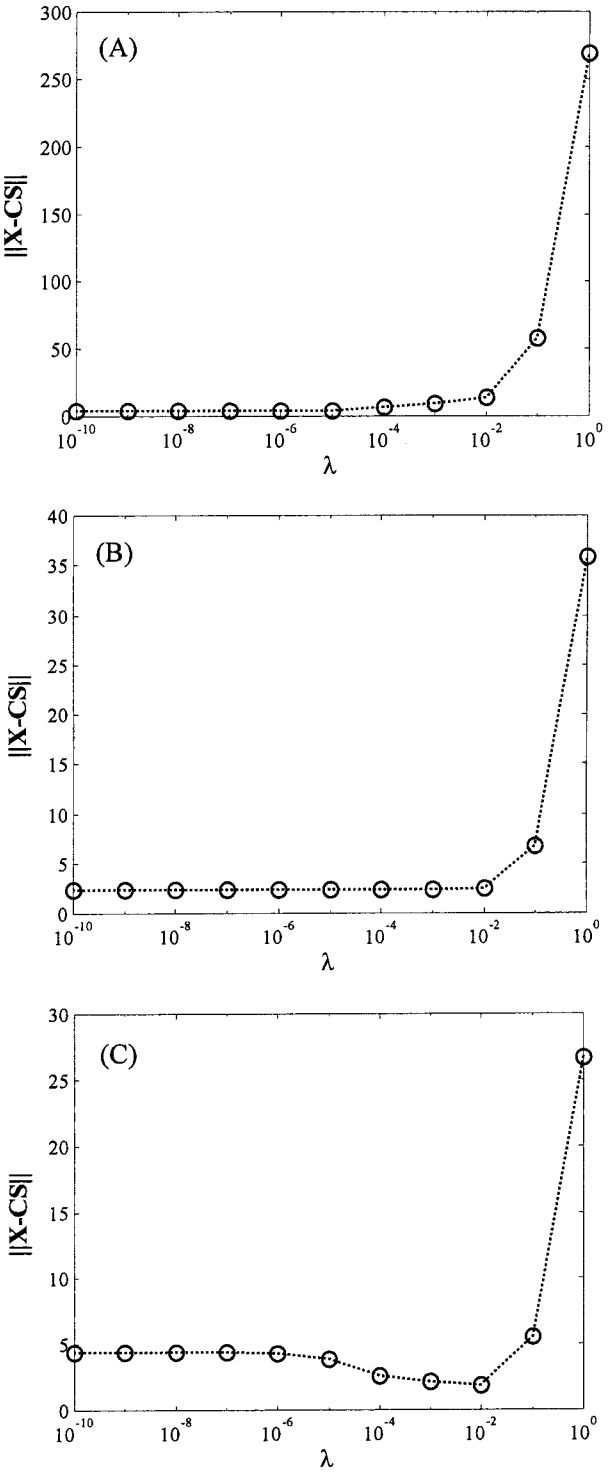

Fig. 8. Frobenius norm of **X-CS**, $||\mathbf{X} - \mathbf{CS}||$, calculated by changing the value of λ from 10^{-10} to 1. (A) $||\mathbf{X} - \mathbf{CS}||$ for data 1. (B) $||\mathbf{X} - \mathbf{CS}||$ for data 2. (C) $||\mathbf{X} - \mathbf{CS}||$ for data 3.

4.6. Theoretical problems of RALS and ICA and future development

If the matrix **C'C** or **SS'** is singular, the regularized term is often useful. In such cases, the regularized parameter λ should be set to a small positive value. In the present study, since the matrix **C'C** or **SS'** is non-singular and the data contain little noise, we failed to clarify the advantage of RALS.

In the present study, we calculated ICA on the assumption that each column of **C** in Eq. (1) is independent of each other. By the theoretical character of ICA, heavily overlapping peaks violate the assumption of independence and make it difficult to resolve the spectrum. Because some metabolites have similar peak positions in the DAD spectrum, it is difficult to apply ICA in cases where each row of **S** in Eq. (1) corresponding to the spectrum is considered to be independent of the others. Indeed, it was impossible to resolve chromatographic peaks for data 1 by ICA assuming that each row of **S** in Eq. (1) is independent of each other. For data 2 and data 3, however, the results obtained on the assumption were similar to those in Fig.7 (E), (F) (data not shown). When ICA was applied to the real dataset, it should be noted to check whether statistical independence is valid for the data. To match with a real dataset, an improved ICA algorithm with non-negativity constraint may be needed. Recently, ICA with non-negativity constraints, non-negative ICA, has been proposed and further development of ICA can be expected in the SMCR context [13].

References

[1] E. Fukusaki, A. Kobayashi, Plant metabolomics: potential for practical operation, J. Biosci. Bioeng. 100 (2005) 347-354.

[2] J. Jiang, Y. Liang, Y. Ozaki, Principles and methodologies in self-modeling curve resolution, Chemom. Intell. Lab. Syst. 71 (2004) 1-12.

[3] E. J. Karjalainen, The spectrum reconstruction problem, use of alternating regression for unexpected spectral components in two-dimensional spectroscopies, Chemom. Intell. Lab. Syst. 7 (1989) 31-38.

[4] P. Jonsson, AI. Johansson, J. Gullberg, J. Trygg, A J, B. Grung, S. Marklund, M. Sjostrom, H. Antti, T. Moritz, High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses, Anal Chem. 77 (2005) 5635-5642.

[5] R. Tauler, Multivariate curve resolution applied to second order data, Chemometr. Intell. Lab. Syst. 30 (1995) 133-146.

[6] A. Hyvarinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley Interscience, New York, 2001.

[7] J. Chen, XZ. Wang, A new approach to near-infrared spectral data analysis using independent component analysis, J. Chem. Inf. Comput. Sci.41 (2001) 992-1001.

[8] T. Katsuda, A. Lababpour, K. Shimahara, S. Katoh, Astaxanthin production by *Haematococcus pluvialis* under illumination with LEDs, Enzyme Microb. Technol. 35, (2004) 81-86.

[9] Y. Hu, W. Shen, W. Yao, D. L. Massart, Using singular value ratio for resolving peaks in HPLC-DAD data sets, Chemom. Intell. Lab. Syst. 77 (2005) 97-103.

[10] I. E. Frank, J. H. Friedman, A statistical view of some chemometrics regression tools, Technometrics 35 (1993) 109-135.

[11] A. Hyvarinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Trans. Neural Netw. 10 (1999) 626-634.

[12] J. P. Yuan, F. Chen, Chromatographic separation and purification of trans-astaxanthin from the extracts of *Haematococcus pluvialis*, J. Agric. Food Chem., 46 (1998) 3371-3375.

[13] M. D. Plumbley, Algorithms for non-negative independent component analysis, IEEE Trans. Neural Netw. 14 (2003) 534-543.

GENERAL CONCLUSION

Application of multivariate analysis to metabolome data for the purpose of visualization and discrimination, regression, and curve resolution is achieved.

Multivariate analysis for regression was studied in part I. RCCA for multivariate regression was investigated and compared with PLS. The optimal number of latent variables in RCCA determined by LOOCV was significantly fewer than in PLS. PLS extracts a number of latent variables that are redundant to the construction of a quality-predictive model because the explanatory variable **X**, which is not closely related to the response variable **y** and has a large variance, affects the model more severely in PLS than in RCCA. A long calculation time is required for LOOCV of RCCA. This problem can be avoided by using kernel CCA with a linear kernel, which results in a reduction of the calculation time of RCCA from 5.9962×10^4 s (about 16.5 h) to 8.1090 s. These results suggest that RCCA as well as PLS is useful for multivariate regression.

Multivariate analysis for visualization and discrimination was studied in part II. We extended some ordinary dimensionality reduction methods, principal component analysis (PCA), partial least squares (PLS), orthonormalized PLS, and regularized Fisher discriminant analysis (RFDA) by introducing the differential penalty of the latent variables with each class. A nonlinear extension to these methods by kernel methods was also proposed as kernel smoothed PCA, PLS, OPLS, and FDA. The effect of smoothness can be found by calculation results. These methods are useful for the data in which observation is in transition with time.

Multivariate analysis for curve resolution was studied in part III. We proposed a regularized version of ALS, known as RALS. We applied it by adding a normalized constraint to a real dataset consisting of the HPLC-DAD data of *H. pluvialis* metabolites. We then used ICA to resolve the HPLC-DAD data. The results suggested that RALS gives different solutions by changing the regularized parameter λ and these curve

resolution methods are useful for peak detection of metabolites in HPLC-DAD data.

PUBLICATION LIST

Part I

Yamamoto, H., Yamaji, H., Fukusaki, E., Ohno, H., Fukuda H. Canonical correlation analysis for multivariate regression and its application to metabolic fingerprinting. Biochemical Engineering Journal (2008), in press

Part II

Yamamoto, H., Yamaji, H., Abe, Y., Harada, K., Danang, W., Fukusaki, E., Ohno, H., Kondo, A., Fukuda, H. Dimensionality reduction by PCA, PLS, OPLS, RFDA with smoothness (in preparation)

Part III

Yamamoto, H., Hada, K., Yamaji, H., Katsuda, T., Ohno, H., Fukuda H.
Application of regularized alternating least squares and independent component analysis to HPLC-DAD data of *Haematococcus pluvialis* metabolites, Biochemical Engineering Journal, 32 (2006) 149-156