Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2012 *Paper* 100

PLS-ROG: Partial least squares with rank order of groups

Hiroyuki Yamamoto*

*Human Metabolome Technologies, Inc., h.yama2396@gmail.com

PLS-ROG: Partial least squares with rank order of groups

Hiroyuki Yamamoto

Abstract

Partial least squares (PLS), which is an unsupervised dimensionality reduction method, has been widely used in metabolomics. PLS can separate score depend on groups in a low dimensional subspace. However, this cannot use the information about rank order of groups. This information is often provided in which concentration of administered drugs to animals is gradually varies. In this study, we proposed partial least squares for rank order of groups (PLS-ROG). PLS-ROG can consider both separation and rank order of groups.

Partial least squares for rank order of groups (PLS-ROG)

Consider a matrix Y, which is a mean-centered dummy matrix of group information whose elements are 0 or 1. The latent variables $\mathbf{t}$ and $\mathbf{s}$ are related to the data matrix $\mathbf{X}$ and the dummy matrix $\mathbf{Y}$ by $\mathbf{t} = \mathbf{X}\mathbf{w_x}$ and $\mathbf{s} = \mathbf{Y}\mathbf{w_y}$. PLS for discrimination [1] is formulated as the optimization problem of maximizing the covariance between the latent explanatory variable $\mathbf{t}$ and the latent response variable $\mathbf{s}$:

$$\begin{aligned}&\max \operatorname{cov}(\mathbf{t},\mathbf{s})\\ &\textit{subject to } \mathbf{w_x}'\mathbf{w_x} = 1,\ \mathbf{w_y}'\mathbf{Y}'\mathbf{Y}\mathbf{w_y} = 1\end{aligned} \qquad (1)$$

The differential penalty of mean of the latent variable $\mathbf{s}$ in each group is added to the constraint condition of PLS. PLS-ROG is formulated as follows:

$$\begin{aligned}&\max \operatorname{cov}(\mathbf{t},\mathbf{s})\\ &\textit{subject to } \mathbf{w}_x'\mathbf{w}_x = 1,\ \mathbf{w_y}'Y'Y\mathbf{w_y} + \kappa\mathbf{w_y}'\mathbf{Y}'\mathbf{P}'\mathbf{D}'\mathbf{D}\mathbf{P}\mathbf{Y}\mathbf{w_y} = 1\end{aligned}$$

This formulation is similar as our previous study [2]. The differential matrix $\mathbf{D}$ and averaging matrix of groups $\mathbf{P}$ are set for the g class classification problems as follows:

$$D = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix},$$

$$P = \begin{bmatrix} 1/n_1 & \cdots & 1/n_1 & \mathbf{0} & & \mathbf{0} & \\ & \mathbf{0} & & \ddots & & \mathbf{0} & \\ & \mathbf{0} & & \mathbf{0} & 1/n_g & \cdots & 1/n_g \end{bmatrix}$$

By using the method of Lagrange multipliers, the problem in eq. (1) can be reformulated as the maximization of

$$\begin{aligned}J = {}&\frac{1}{n-1}\mathbf{w_x}'\mathbf{X}'\mathbf{Y}\mathbf{w_y} + \lambda_x(1 - \mathbf{w_x}'\mathbf{w_x})\\ &+ \lambda_y(1 - \mathbf{w_y}'\mathbf{Y}'\mathbf{Y}\mathbf{w_y} - \kappa\mathbf{w_y}'\mathbf{Y}'\mathbf{P}'\mathbf{D}'\mathbf{D}\mathbf{P}\mathbf{Y}\mathbf{w_y})\end{aligned} \qquad (2)$$

Partial differentiation of eq. (2) with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$, followed by a transformation, yields the following two equations.

$$\frac{1}{n-1}\mathbf{X'Yw_y} = 2\lambda_x \mathbf{w}_x \qquad (3)$$

$$\frac{1}{n-1}\mathbf{Y'Xw_x} = 2\lambda_y(\mathbf{Y'Y} + \kappa\mathbf{Y'P'D'DPY})\mathbf{w_y} \qquad (4)$$

Eqs. (3) and (4) can be rewritten as the eigenvalue problems

$$\frac{1}{(n-1)^2}\mathbf{X'Y}(\mathbf{Y'Y} + \kappa\mathbf{Y'P'D'DPY})^{-1}\mathbf{Y'Xw}_x = \lambda\mathbf{w}_x \qquad (5)$$

$$\frac{1}{(n-1)^2}\mathbf{Y'XX'Yw}_y = \lambda(\mathbf{Y'Y} + \kappa\mathbf{Y'P'D'DPY})\mathbf{w}_y. \qquad (6)$$

where $\lambda = 4\lambda_x\lambda_y$. These eigenvalue problems can be computed by using singular value decomposition.

Factor loading in PLS-ROG

We now describe the statistical properties of the eigenvector $\mathbf{w}_x$ that can be used for factor loading in PLS-ROG. The correlation coefficient between the latent response variables $\mathbf{s}$ and the p-th explanatory variable $\mathbf{x_p}$ can be written as

$$\text{corr}\,(\mathbf{s}, \mathbf{x_p}) = \frac{\mathbf{s'x_p}/n-1}{\sqrt{\text{var}(\mathbf{s})}\sqrt{\text{var}(\mathbf{x_p})}} \qquad (7)$$

Substituting $\mathbf{s} = \mathbf{Yw}_y$ and $\mathbf{x_p} = \mathbf{Xc}$, where $\mathbf{c}$ the column vector in which the p-th element is 1 and the others are 0, yields

$$\text{corr}(\mathbf{s}, \mathbf{x_p}) = \frac{\mathbf{w_y'Y'Xc}/n-1}{\sqrt{\text{var}(\mathbf{s})}\sqrt{\text{var}(\mathbf{x_p})}} \qquad (8)$$

Transposing eq. (3) gives $\mathbf{w_y'Y'X}/n-1 = 2\lambda_x\mathbf{w}_x'$ which can be substituted in eq. (8), giving

$$\text{corr}(\mathbf{s}, \mathbf{x_p}) = \frac{2\lambda_x\mathbf{w}_x'\mathbf{c}}{\sqrt{\text{var}(\mathbf{s})}\sqrt{\text{var}(\mathbf{x}_p)}} \qquad (9)$$

Now

$$\mathrm{var}(\mathbf{s}) = \frac{1}{\mathrm{n}-1}\mathbf{w}_y\mathbf{'Y'Yw_y} \quad (10)$$
,

which can be substituted into eq. (9), giving

$$\begin{aligned}\mathrm{corr}(\mathbf{s},\mathbf{x_p}) &= \frac{2\lambda_x \mathbf{w}_x\mathbf{'c}}{\sqrt{\mathrm{var}(\mathbf{s})}\sqrt{\mathrm{var}(\mathbf{x}_p)}} = \frac{2\lambda_\mathbf{x} w_{x,p}}{\sqrt{\mathbf{w_y'Y'Yw_y}/(n-1)}\,\sigma_p} \\ &= \frac{\sqrt{(n-1)\lambda}\,w_{x,p}}{\sqrt{\mathbf{w_y'Y'Yw_y}}\,\sigma_p}\end{aligned} \quad (11)$$
.

The scalar $\mathbf{w_y'Y'Yw_y}$ can be assumed to be constant because it does not depend on the p-th variable $\mathbf{x_p}$. With autoscaling of data, the p-th component of the eigenvector $\mathbf{w}_x$ is proportional to the correlation coefficient between the latent response variable $\mathbf{s}$ and the p-th explanatory variable $\mathbf{x_p}$. For this reason, factor loadings in PLS can be defined by correlation coefficient in eq. (11). Using this definition we can perform a statistical test for factor loading.

Reference

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

[2] H.Yamamoto, H.Yamaji, Y.Abe, K.Harada, D.Waluyo, E.Fukusaki, A.Kondo, H. Ohno, H.Fukuda, Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables, Chemom. Intell. Lab. Syst., 98 (2009) 136-142