Vol.2015-BIO-44 No.4
2015/12/7

カーネル Partial least squares-rank order of groups －複数臓器由来メタボロームデータの統合解析 －

山本博之†

概要：我々は以前，教師あり次元削減手法の 1 つである Partial least squares (PLS)に平滑化の罰則項を加え，群に順序がある時に適した PLS-ROG (Partial least squares-rank order of groups)を提案した．PLS-ROG は，カーネル法を用いてカーネル PLS-ROG へと拡張することが出来る．本報告では，カーネル PLS-ROG を用いて，同一個体から採取した様々な臓器・生体液のメタボロームデータを統合し，解析を行った．

キーワード：カーネル法，グループラッソ，Partial least squares，メタボロミクス

Kernel Partial least squares-rank order of groups －Integrated analysis of multiple region-derived metabolome data－

HIROYUKI YAMAMOTO†

Abstract: We proposed PLS-ROG (Partial least squares-rank order of groups) that is suited for rank order of groups. We also extended PLS-ROG to kernel PLS-ROG by using kernel method. In the present study, we analyzed multiple region-derived metabolome data obtained from the same individuals by using kernel PLS-ROG.

Keywords: Kernel method, Group lasso, Partial least squares, Metabolomics

1. はじめに

生体内の代謝物を包括的に解析する研究分野であるメタボロミクスにおいて，教師あり次元削減法 1 つである Partial least squares(PLS)は，視覚化・回帰・判別モデルの構築など，様々な目的で用いられている[1, 2]．我々は，PLS に平滑化の罰則項を加え，群の順序を考慮した PLS である PLS-ROG (Partial least squares-rank order of groups)を提案し，実際のメタボロームデータへ適用してきた[3]．本報告では，さらにカーネル法を用いてカーネル PLS-ROG へと拡張し，同一個体から採取した様々な臓器，生体液のメタボロームデータを統合して解析を行った．

2. 理論

2-1. PLS-ROG

PLS-ROG は，以下の条件式

$$\max \operatorname{cov}(\mathbf{t},\mathbf{s}) \quad subject\ to\ \mathbf{w}_x'\mathbf{w}_x=1,(1-\kappa)\mathbf{w}_y'\mathbf{w}_y+\kappa\mathbf{w}_y'\mathbf{Y}'\mathbf{P}'\mathbf{D}'\mathbf{DPY}\mathbf{w}_y=1 \qquad \text{式 (1)}$$

に基づいた最適化問題として定式化される．**X** を n(サンプル)×p(代謝物)の行列，**Y** を群情報のダミー行列[1]とする．説明変数と目的変数の合成変数 **t**, **s** は，それぞれ **X** の重みベクトル $\mathbf{w_x}$, **Y** の重みベクトル $\mathbf{w_y}$ による線形結合 $\mathbf{t}=\mathbf{Xw_x}$, $\mathbf{s}=\mathbf{Yw_y}$ で表される．ただし，行列 **D** と **P** は以下の行列

$$\mathbf{D}=\begin{bmatrix}-1&1&0&0&\cdots&0&0\\0&0&-1&1&\cdots&0&0\\\vdots&\vdots&\vdots&\vdots&\ddots&\vdots&\vdots\\0&0&0&0&\cdots&-1&1\end{bmatrix} \quad \mathbf{P}=\begin{bmatrix}1/n_1\cdots1/n_1&0&0\\0&\ddots&0\\0&0&1/n_g\cdots1/n_g\end{bmatrix}$$

で表される．ラグランジュ乗数法より，条件式(1)は次式のように書ける．

$$J=\frac{1}{n-1}\mathbf{w_x}'\mathbf{X}'\mathbf{Y}\mathbf{w_y}+\lambda_x(1-\mathbf{w_x}'\mathbf{w_x})+\lambda_y\left\{1-(1-\kappa)\mathbf{w_y}'\mathbf{w_y}-\kappa\mathbf{w_y}'\mathbf{Y}'\mathbf{P}'\mathbf{D}'\mathbf{DPY}\mathbf{w_y}\right\}$$

$\mathbf{w_x}$ と $\mathbf{w_y}$ でそれぞれ偏微分し整理すると，最終的に次の一般化固有値問題で書ける．

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

$$\frac{1}{(n-1)^2}\mathbf{Y}'\mathbf{X}\mathbf{X}'\mathbf{Y}\mathbf{w}_y=\lambda\left\{(1-\kappa)\mathbf{I}+\kappa\mathbf{Y}'\mathbf{P}'\mathbf{D}'\mathbf{DPY}\right\}\mathbf{w}_y \qquad \text{式 (2)}$$

次に，諏訪らの主成分分析（PCA）の場合[4]に従って，式(1)の制約条件 $\mathbf{w_x}'\mathbf{w_x}=1$ を，グループラッソの罰則項 $\|\mathbf{w}\|_2 \leqq 1$ で置き換える．相加相乗平均の関係より，$\mathbf{w_x}'\mathbf{w_x}=1$ を $(\|\mathbf{w_{xm}}\|^2/\beta_m+\beta_m)/2=1$ で置き換えることが出来る[4]．この制約条件下では，β_m は PCA[4]の場合と同じく，

$$\beta^{new}=\sqrt{\beta_m^{old}}\left\|\tilde{\mathbf{w}}_{\mathbf{xm}}\right\| \qquad \text{式 (3)}$$

によって更新することにより計算できる．ここで，$\tilde{\mathbf{w}}_{\mathbf{xm}}=\mathbf{w_{xm}}/(\beta_m)^{1/2}$ とおいた．$\mathbf{w_x}$とβ_mの計算は，式 (2) と式 (3) の繰り返し計算によって行う．

† ヒューマン・メタボローム・テクノロジーズ(株)
研究開発本部　メタボロミクス基盤研究部
Human Metabolome Technologies Inc.
Research & Development, Metabolomics Research

2-2. カーネル PLS-ROG

PLS-ROG をカーネル法を用いて非線形手法へと拡張したカーネル PLS-ROG は，以下の条件式

$$\max \operatorname{cov}(\mathbf{t},\mathbf{s}) \quad subject\ to\ \boldsymbol{\alpha}_x'K\boldsymbol{\alpha}_x=1,(1-\kappa)\mathbf{w}_y'\mathbf{w}_y+\kappa\mathbf{w}_y'\mathbf{Y}'\mathbf{P}'\mathbf{D}'\mathbf{DPY}\mathbf{w}_y=1 \qquad \text{式（4）}$$

に基づいた最適化問題として定式化される．スコア $\mathbf{t}$, $\mathbf{s}$ は，それぞれ $\mathbf{K}$ の重みベクトル $\boldsymbol{\alpha}_x$，$\mathbf{Y}$ の重みベクトル $\mathbf{w}_y$ による線形結合 $\mathbf{t}=\mathbf{K}\boldsymbol{\alpha}_x$, $\mathbf{s}=\mathbf{Y}\mathbf{w}_y$ で表される．カーネル行列 $\mathbf{K}$ は，データ行列 $\mathbf{X}$ の非線形変換 $\boldsymbol{\Phi}(\mathbf{X})$により，$\mathbf{K}=\boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\Phi}(\mathbf{X})'$ と書ける．

ラグランジュ乗数法より，条件式(4)は次式で書ける．

$$J=\frac{1}{n-1}\boldsymbol{\alpha}_x'\mathbf{KY}\mathbf{w}_y+\lambda_x(1-\boldsymbol{\alpha}_x'\mathbf{K}\boldsymbol{\alpha}_x)+\lambda_y\left\{1-(1-\kappa)\mathbf{w}_y'\mathbf{w}_y-\kappa\mathbf{w}_y'\mathbf{Y}'\mathbf{P}'\mathbf{D}'\mathbf{DPY}\mathbf{w}_y\right\}$$

$\boldsymbol{\alpha}_x$ と $\mathbf{w}_y$ でそれぞれ偏微分し整理すると，最終的に次の一般化固有値問題で書ける．

$$\frac{1}{(n-1)^2}\mathbf{KY}\left\{(1-\kappa)\mathbf{I}+\kappa\mathbf{Y}'\mathbf{P}'\mathbf{D}'\mathbf{DPY}\right\}^{-1}\mathbf{Y}'\mathbf{K}\boldsymbol{\alpha}_x=\lambda\mathbf{K}\boldsymbol{\alpha}_x$$

$$\frac{1}{(n-1)^2}\mathbf{Y}'\mathbf{KY}\mathbf{w}_y=\lambda\left\{(1-\kappa)\mathbf{I}+\kappa\mathbf{Y}'\mathbf{P}'\mathbf{D}'\mathbf{DPY}\right\}\mathbf{w}_y \qquad \text{式（5）}$$

ここで，各臓器由来のデータからそれぞれカーネル行列 $\mathbf{K}_m$ を計算し，係数 β_m による線形結合

$$\mathbf{K}=\sum_{m=1}^{M}\beta_m\mathbf{K}_m$$

をカーネル行列 $\mathbf{K}$ とする．β_m は線形手法での，例えば式(3)における β_m そのものであり，文献[4]にある通り，β_m は各臓器由来のカーネル行列に対する重要度とみなすことが出来る．

式(4)の制約条件における $\boldsymbol{\alpha}_x'\ \mathbf{K}\boldsymbol{\alpha}_x=1$ を，グループラッソの制約条件 $\|\mathbf{w}\|_2=\|\boldsymbol{\Phi}(\mathbf{X})'\boldsymbol{\alpha}_x\|_2\leqq 1$ に置き換え，さらに相加相乗平均の関係を用いれば，グループラッソの制約条件を用いたカーネル PLS-ROG は，式(5)と式(6)

$$\beta_m^{new}=\beta_m^{old}\sqrt{\boldsymbol{\alpha}'\mathbf{K}_m\boldsymbol{\alpha}} \qquad \text{式（6）}$$

の繰り返しによって計算することが出来る．

3. データ

Wild type のウサギ(n=3)，高脂血症モデルウサギ WHHL ウサギ(n=3)，WHHL ウサギにスタチンを投与した時(n=3)の肝臓・心臓・脳・血漿サンプルについて，キャピラリー電気泳動-飛行時間型質量分析計を用いて測定し，得られたメタボロームデータを解析に用いた[5]．データはそれぞれ代謝物毎に平均 0，分散 1 に autoscaling を行った．カーネル関数は，線形カーネルを用いた．

4. 解析結果

大賀ら[5]は，野生型・WHHL ウサギ+薬剤投与・WHHL ウサギの順で，上昇もしくは低下するパターンを示す代謝物に着目し，*N*,*N*-Dimethylglycine と Betaine，またプリン代謝の代謝中間体について生物学的な議論を行っている．

本研究では，データをこの並びの順とし，カーネル PLS(κ=0)とカーネル PLS-ROG(κ=0.5)を用いて解析を行った．カーネル PLS では，第 1 成分または第 2 成分で，この順序のパターンを示すスコアは得られていない(図 1 左)．一方で，カーネル PLS-ROG(図 1 右)では，第 1 成分のスコアで，理想的なパターンを示していることがわかる．

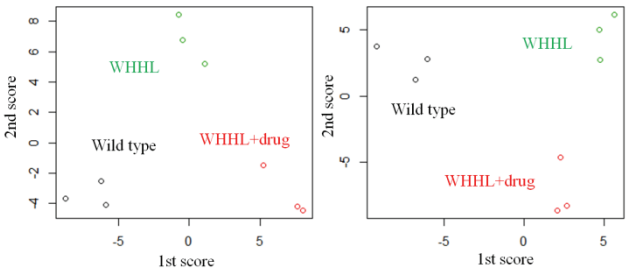

図 1. カーネル PLS(左)，カーネル PLS-ROG (右)の結果

次に，第 1 成分のスコアと各代謝物レベルとの相関係数とその仮説検定を行い，有意な代謝物を選択した．結果は，肝臓と心臓においてグリシン生合成経路の代謝中間体である *N*,*N*-Dimethylglycine（肝臓：R=0.8010，p=0.0095，心臓：R=0.8216，p=0.0066）と Betaine（肝臓：R=0.7266，p=0.0266，心臓：R=0.7739，p=0.0144）で第 1 成分のスコアと有意に正の相関が認められた．また肝臓において，プリン代謝の代謝中間体である Uric acid（R=0.6746，p=0.0462）で有意に正の相関を示し，Hypoxanthine（R=−0.8425，p=0.0043），Inosine（R=−0.8228，p=0.0065），Adenosine（R=−0.7721，p=0.0148）で有意に負の相関が認められ，既存の報告[4]とその傾向は一致していた．

5. おわりに

カーネル PLS-ROG を用いて，複数の臓器と血漿サンプルのメタボロームデータを統合的に解析した．今後はさらに，他のオミクスデータの統合解析への適用を検討していく予定である．

参考文献

1) M.Barker and W.Rayens, "Partial least squares for discrimination", J.Chemom., 17(3), 166-173(2003)
2) S.Wold, M.Sjostrom and L.Eriksson, "PLS-regression: a basic tool of chemometrics", Chem.Int.Lab.Sys., 58(2), 109-130(2001)
3) H.Yamamoto, “PLS-ROG: Partial least squares with rank order of groups.” (in preparation).
4) 諏訪恭平，冨岡亮太，矢入健久，鹿島久嗣，”複数情報源に対する主成分分析”，電子情報通信学会技術研究報告. IBISML，情報論的学習理論と機械学習 110(476), 147-152(2011)
5) T.Ooga, H.Sato, A.Nagashima, K.Sasaki, M.Tomita, T.Soga and Ohashi Y., “Metabolomic anatomy of an animal model revealing homeostatic imbalances in dyslipidaemia.” Mol. Biosyst. 7(4), 1217-1223 (2011).