# One-sided Kernel Principal Component Analysis

**Hiroyuki Yamamoto**
h.yama2396@gmail.com

Japan Computational Mass Spectrometry (JCompMS) group

**Abstract**

Principal component analysis (PCA) is widely used in omics research, such as metabolomics. Kernel principal component analysis (KPCA) is also widely used in machine learning because it can compute various nonlinear PCA depending on the flexible setting of the kernel function, but it is rarely used in omics research. One of the reasons why KPCA has not been used for omics data analysis is that it cannot directly calculate principal component loadings to select important variables such as metabolites. In this study, we propose one-sided KPCA that can directly compute and use principal component loadings to select important variables. Furthermore, one-sided KPCA can utilize similarity data in place of a kernel matrix, enabling it to be used for integrative analysis between different types of data.

**Keyword** principal component analysis, kernel method, metabolomics

# 片側カーネル主成分分析

質量分析インフォマティクス研究会　山本　博之 h.yama2396@gmail.com

**Abstract** 主成分分析は，メタボロミクスをはじめとするオミックス研究において広く用いられている．カーネル主成分分析は，カーネル関数の設定次第で様々な非線形の主成分分析を計算出来ることから，機械学習においては広く用いられているが，オミックス研究においてはほとんど利用されていない．カーネル主成分分析がオミックスデータ解析に用いられてこなかったのは，カーネル主成分分析では主成分負荷量を直接計算することが出来ず，重要な変数を選び出せないことが理由の一つであると考えられる．そこで本研究では，主成分負荷量を計算可能なカーネル主成分分析として，片側カーネル主成分分析を提案する．また片側カーネル主成分分析は，類似度データをカーネル行列の代わりに利用することで，異なるデータ間の統合解析にも利用できる．

## 1. はじめに

サンプル中の代謝物，タンパク質，遺伝子発現などを網羅的に解析するオミックス研究では，得られたデータを可視化し，関連する代謝物，タンパク質，遺伝子といった変数を特定するために，主成分分析が用いられる．主成分分析では，主成分スコアを用いてデータを可視化した後，表現型と関連する主成分スコアを選び出し，主成分スコアと相関の高い変数を主成分負荷量を用いて選ぶ[1]．この選ばれた変数を用いて生物学的な推論を行う．

非線形の主成分分析であるカーネル主成分分析[2]は，データに含まれる変数の非線形変換を明示的に設定する代わりに，カーネル関数を利用することで非線形の主成分分析を行う方法である．カーネル主成分分析は，カーネル関数を変えるだけで，様々な非線形主成分分析を統一的に表すことが出来る一方で，主成分負荷量を用いて重要な変数を選ぶことが出来ない．表現型と関連するスコアと相関の高い変数を選ぶことが出来なければ，解析結果を生物学的な推論へ繋げることが難しく，この点がカーネル主成分分析がオミックス研究で用いられない理由の一つであると考えられる．

そこで本研究では，主成分負荷量を用いて重要な変数を選ぶことが出来るカーネル主成分分析として，片側カーネル主成分分析を提案する．片側カーネル主成分負荷量は，カーネルのスコアと各変数との相関係数で定義出来ることから，統計的仮説検定を用いて統計的に有意な変数を選ぶことが出来る．実際のオミックス研究への適用例として，COVID-19 のメタボロームデータと，マウス糞便のメタボロームデータと 16SrRNA の統合解析に適用した結果について紹介する．

## 2. 理論

### 2-1. 片側カーネル主成分分析

片側カーネル主成分分析は，以下のような線形のスコア **t** とカーネルのスコア **u** の共分散最大化を考える．

$$\max \quad cov(\mathbf{t},\mathbf{u})$$
$$\mathbf{w_x'w_x} = 1, \mathbf{w_z'w_z} = 1$$

ここで，$\mathbf{t}=\mathbf{Xw_x}$，$\mathbf{u}=\mathbf{\Phi w_z}$とする．また $\mathbf{\Phi}$ はデータ行列 **X** を非線形変換したものとする．次にラグランジュ乗数法より

$$J = \frac{1}{N}\mathbf{w_x'X'\Phi w_z} + \lambda_x(1 - \mathbf{w_x'w_x}) + \lambda_z(1 - \mathbf{w_z'w_z})$$

J を $\mathbf{w_x}$, $\mathbf{w_z}$それぞれで偏微分すると

$$\frac{\partial J}{\partial \mathbf{w_x}} = \frac{1}{N}\mathbf{X'\Phi w_z} - 2\lambda_x\mathbf{w_x} = 0$$
$$\frac{\partial J}{\partial \mathbf{w_z}} = \frac{1}{N}\mathbf{\Phi'X w_x} - 2\lambda_z\mathbf{w_z} = 0$$

$\mathbf{w_x}$, $\mathbf{w_z}$それぞれについて整理すると

$$\mathbf{X'\Phi\Phi'Xw_x} = 4N^2\lambda_x\lambda_z\mathbf{w_x}$$
$$\mathbf{\Phi'XX'\Phi w_z} = 4N^2\lambda_x\lambda_z\mathbf{w_z}$$

$\lambda=4N^2\lambda_x\lambda_y$とおくと

$$\mathbf{X'\Phi\Phi'Xw_x} = \lambda\mathbf{w_x}$$
$$\mathbf{\Phi'XX'\Phi w_z} = \lambda\mathbf{w_z}$$

ここで $\mathbf{\Phi'XX'\Phi w_z} = \lambda\mathbf{w_z}$について，左から$\mathbf{\Phi}$を掛け，また $\mathbf{w_z}= \mathbf{\Phi'}\boldsymbol{\alpha_z}$とおくと

$$\mathbf{\Phi\Phi'XX'\Phi\Phi'}\boldsymbol{\alpha_z} = \lambda\mathbf{\Phi\Phi'}\boldsymbol{\alpha_z}$$

カーネル行列 $\mathbf{K}=\mathbf{\Phi\Phi'}$を代入して

$$\mathbf{X'KXw_x} = \lambda\mathbf{w_x}$$
$$\mathbf{K'XX'K}\boldsymbol{\alpha_z} = \lambda\mathbf{K}\boldsymbol{\alpha_z}$$

となり，最終的に $\mathbf{w_x}$についての固有値問題と $\boldsymbol{\alpha_z}$についての一般化固有値問題で書ける．

また片側カーネル主成分分析は，目的変数にデータのカーネル行列を設定した時の PLS[3]であると解釈することが出来る．

### 2-2. 片側カーネル主成分負荷量の統計的な性質

スコア **u** と **X** の p 番目の変数 $\mathbf{x_p}$の相関係数は以下のように書ける．

$$cor(\boldsymbol{u},\boldsymbol{x_p}) = \frac{cov(\boldsymbol{u},\boldsymbol{x_p})}{\sqrt{var(\boldsymbol{u})}\sqrt{var(\boldsymbol{x_p})}} = \frac{\frac{1}{N}\mathbf{w_z'\Phi'Xc}}{\sqrt{\frac{1}{N}\mathbf{w_z\Phi'\Phi w_z}}}$$

ここでデータは各変数に対して平均 0，分散 1 に autoscaling されているとする．$\frac{1}{N}\mathbf{X'\Phi w_z} = 2\lambda_x\mathbf{w_x}$の両辺を転置したものを上式に代入して，

$$cor(\boldsymbol{u},\boldsymbol{x_p}) = \frac{2\lambda_x\mathbf{w_x c}}{\sqrt{\frac{1}{N}\mathbf{w_z'\Phi'\Phi w_z}}} = \frac{2\lambda_x\mathbf{w_{x,p}}}{\sqrt{\frac{1}{N}\mathbf{w_z'\Phi'\Phi w_z}}}$$

$\mathbf{w_z}= \mathbf{\Phi'}\boldsymbol{\alpha_z}$を代入して

$$cor(\boldsymbol{u}, \boldsymbol{x_p}) = \frac{2\lambda_x \mathbf{w_{x,p}}}{\sqrt{\frac{1}{N}\boldsymbol{\alpha_z}'\boldsymbol{\Phi}\boldsymbol{\Phi}'\boldsymbol{\Phi}\boldsymbol{\Phi}'\boldsymbol{\alpha_z}}} = \frac{2\lambda_x \mathbf{w_{x,p}}}{\sqrt{\frac{1}{N}\boldsymbol{\alpha_z}'\mathbf{K^2}\boldsymbol{\alpha_z}}}$$

ここで, $\lambda_x=(1/2)\mathrm{cov}(\mathbf{t},\mathbf{u})$である. また上式の分母は p の影響を受けないので, $w_x$ は $\mathbf{u}$ と $\mathbf{x_p}$ の相関係数に比例することがわかる.

以上より, 片側カーネル主成分負荷量はスコア u と元のデータの各変数の相関係数として定義され, また相関係数の統計的仮説検定を用いることで統計的に有意な変数を選ぶことが出来る.

## 3. 結果と考察

### 3-1. COVID-19 メタボロームデータ

COVID-19 の血清サンプルのメタボロームデータ[4]に対して主成分分析を行った結果を図 1 に示す.

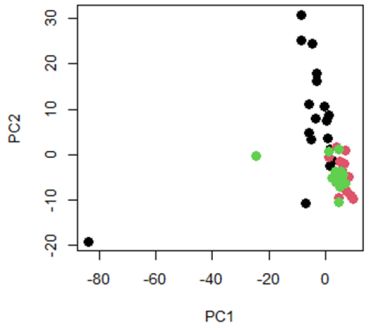


図 1. 主成分分析の結果

●健常群, COVID-19 ●軽症群, ●重症群

主成分分析の結果より, PC2 で健常群と COVID-19 群との違いは確認できるものの, 外れ値の影響が大きいことがわかる. 次にカーネル関数に Hyperbolic tangent kernel を用いた時の片側カーネル主成分分析の結果を図 2 に示す.

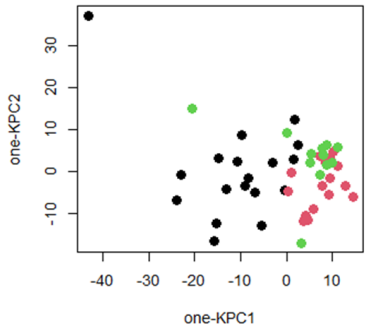


図 2(A). 片側カーネル主成分分析(線形)の結果

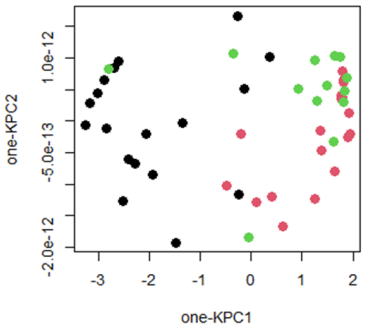


図 2(B). 片側カーネル主成分分析(カーネル)の結果

この結果より, 片側カーネル PC1 では健常群と COVID-19 群の違い, PC2 で COVID-19 軽症群と重傷群の違いが確認できている.

次に, Hyperbolic tangent kernel を用いた時のカーネル主成分分析の結果を図 3 に示す.

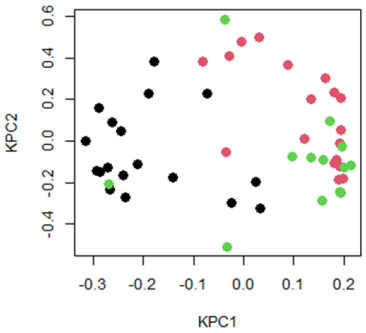


図 3. カーネル主成分分析の結果

カーネル主成分分析の結果より, 図 2(B)の片側カーネル主成分分析の結果と同様の結果が得られることが確認できた.

片側カーネル主成分分析は, 主成分分析とカーネル主成分分析の中間のような結果が得られ, カーネル主成分分析の非線形の良さと, スコアと関連する変数を主成分負荷量から選ぶことが出来る両方の良さを兼ね備えた方法であるといえる.

最後に, 健常群と COVID-19 群の違いが確認された第 1 片側カーネル主成分負荷量について, 相関係数の絶対値が 0.7 以上である代謝物が 19 物質, $p<0.05$ で有意な代謝物が 485 物質、BH 法による $q<0.05$ で有意な代謝物が 378 物質, それぞれ確認された.

### 3-2. 菌叢とメタボロームデータの統合解析

前節では, データそのものと, 非線形変換を行ったカーネル行列(サンプル同士の類似度行列)の組み合わせに対して片側カーネル主成分分析を適用した. これに限らず, 一方がサンプル×変数の行列データ, もう一方がサンプル×サンプルのサンプル同士の類似度行列の組み合に対して, 片側カーネル主成分分析は適用可能である.

例えば 16SrRNA によって得られた weighted unifrac 距離に基づいて計算されるサンプル間の類似度行列と，サンプル×変数の行列データであるメタボロームデータの統合解析に片側カーネル主成分分析を適用した例を紹介する．

マウス糞便の菌叢データとメタボロームデータ[6]の統合解析に片側カーネル主成分分析を適用した．ここで菌叢データとして，16SrRNA の weighted unifrac 解析の結果から距離行列を算出し，その逆数を取った類似度行列を用いた．図 4(A)はメタボロームデータ，図 4(B)は菌叢データの片側カーネル主成分スコアプロットを示している．

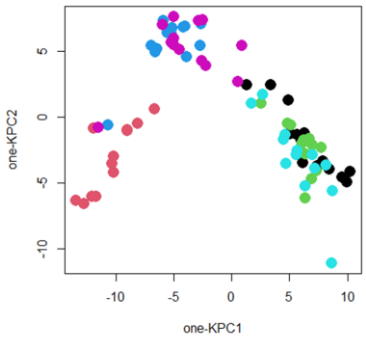


図 4(A)．片側カーネル主成分分析の結果

メタボロームデータ

●●● 3 種類の通常飼育マウス

●●● 3 種類のマウスに抗生物質を投与

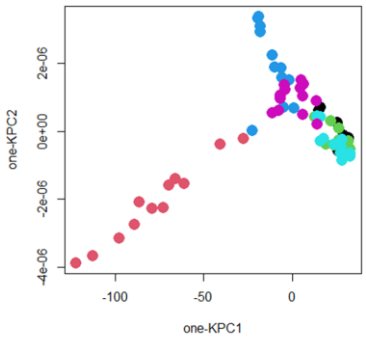


図 4(B)．片側カーネル主成分分析の結果

菌叢データ

この結果より，メタボロームデータ，菌叢データいずれも第 1 成分で抗生物質投与の有無による違いが確認出来たことから，片側カーネル主成分分析を用いることで，これら 2 つのデータセット間の関係を抽出できた．

また通常飼育と抗生物質投与の違いが確認された第 1 片側カーネル主成分負荷量を用いて重要な代謝物を確認した結果，相関係数の絶対値が 0.7 以上である代謝物が 11 物質，$p<0.05$ で有意な代謝物が 135 物質、BH 法による $q<0.05$ で有意な代謝物が 129 物質それぞれ確認された．

## 4. おわりに

本研究では，主成分負荷量が計算可能でかつカーネル法による非線形の主成分分析である片側カーネル主成分分析を提案した．実際に，COVID-19 のメタボロームデータに適用した結果，線形の主成分分析では得られなかった群間差が，片側カーネル主成分分析とカーネル主成分分析で確認できた．さらに片側カーネル主成分負荷量の結果より，健常群と COVID-19 群で差のある代謝物を統計的な基準に基づいて選び出すことが出来た．さらに本手法をマウス糞便の菌叢データとメタボロームデータの統合解析にも適用し，2 つのデータセット間の関連を確認できると共に，関連する重要な代謝物を選び出すことが出来ることが確認された．また本研究で用いた R プログラムは，loadings パッケージとして CRAN[5]で公開している．

## 参考文献


[1] Yamamoto, H. et al. BMC Bioinformatics 15, 51 (2014).
[2] Scholkopf, B. et al. Neural Computation. 10 (5): 1299 (1998).
[3] Yamamoto, H. Journal of Chemometrics. 31(3) e2883 (2017).
[4] Shen B. Cell. 182, 59-72. e15 (2020).
[5] https://cran.r-project.org/web/packages/loadings
[6] Wakita, Y. et al. BMC microbiology, 18:188 (2018)