多変量解析を用いた メタボロームデータ解析

Multivariate Analysis Approach
for Metabolome Data Analysis

4.1　メタボロミクスにおける多変量解析の役割

メタボロミクスにおいて、多変量解析はデータの視覚化、または回帰・判別の予測モデルの構築のために用いられている。多変量解析の手法としてよく知られ、またメタボロミクスで比較的よく用いられる方法として、主成分分析 *(Principal Component Analysis, PCA)[50] または Partial least squares**(PLS)[51] が挙げられる。主成分分析と PLS の違いは、その計算に群情報を用いるか否かであり、前者は教師なし、後者は教師あり次元削減法 *** として区別される。多変量解析の理論的詳細については、様々な書籍 (例えば文献 [52]) に説明されていることから、ここでは省略する。本章では、メタボロミクス研究において、多変量解析、その中でも特に主成分分析と PLS がどのように用いられているか、またその結果をどのようにして生物学的な解釈へと繋げていくか、について説明する。

実際にメタボロミクス研究で、主成分分析がどのように用いられているかを知るために、次のような文献調査を行った。PubMed で、"Metabolomics principal component analysis" をキーワードに検索したところ、244 の文献がヒットした (2010 年 8 月 10 日現在)。その

*　メタボロームデータの合成変数 (= 主成分スコア) の分散最大化を基準とした教師なし次元削減法。メタボロミクスだけでなく、様々な分野で用いられている。

**　説明変数であるメタボロームデータの合成変数と、目的変数の合成変数の共分散最大化を基準とした教師あり次元削減法。派生した方法として、PLS 回帰 , PLS-DA, OPLS 等がある。

***　多変量解析を用いて高次元のデータを2もしくは3次元で表現する方法一般を、特に機械学習のコミュニティでは次元削減法と呼ぶ。主成分分析はデータだけを用いるので教師なし次元削減法、一方 PLS は教師データとして群情報を用いることから、教師あり次元削減法と呼ばれる。

中で、Free Article である 52 の論文に対して、多変量解析の適用法を整理した。その結果を次に示す。

まず、メタボロームデータの 2 または 3 次元での視覚化表現である主成分スコアのプロットを用い、そこから外れ値の確認やサンプルの集合であるクラスターを主観的に判断する[53 54 57 58 61 63 75 81 83 87]、さらに主成分スコアのプロットから、興味あるパターンに対応する主成分軸を選択し、その主成分軸に対応する因子負荷量から特定の物質に着目することで、その後の生物学的な解釈に活かしているもの[59 60 64 65 66 67 68 70 71 72 77 78 79 80 84 85 86 88]、主成分分析の結果だけでなく、PLS による視覚化と因子負荷量からの結果の解釈を行っているもの[55 56 62 65 69 73 74 76 79 80 82 84]、回帰・判別の予測モデルの構築を主な目的としているもの、その他に分けられる。

以上を整理すると、多変量解析を用いてメタボロームデータを 2 もしくは 3 次元に視覚化した後、因子負荷量から特定の物質に着目し、さらなる生物学的解釈へと結び付ける。またその数は少ないものの、メタボロームデータを用いた回帰・判別の予測モデルの構築を狙った研究、その他、におおまかに分類される。そこで本稿では、主に主成分分析と因子負荷量から生物学的な解釈を行う方法について述べる。回帰・判別の予測モデルの構築については、本稿の最後で簡単に説明する。

4.2　主成分分析の結果の見方 : スコアプロットと因子負荷量

前節で説明したように、主成分分析はメタボロミクス研究において、2 もしくは 3 次元の主成分スコア (図 4.1(左)) を観察し、興味ある主成分軸に関与する代謝物質を因子負荷量 (図 4.1(右)) から探し出し、さらなる生物学的解釈へと結び付けるために用いられる。

因子負荷量は、主成分スコアと各物質の相関係数で定義される[89 90 91 92 93]ので、相関係数が正に大きいものは対応する主成分スコアと同じような傾向を示す物質であり、負に大きいものは、主成分スコアとは逆の傾向を示す物質となる。

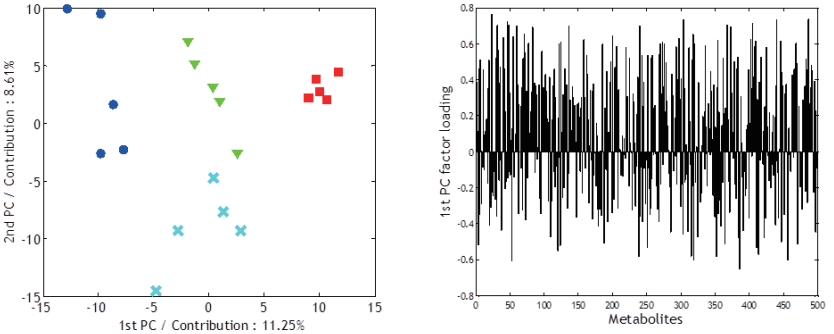

図 4.1　主成分スコア (左) と , 第一主成分の因子負荷量 (右) の結果例

因子負荷量は、主成分スコアと各物質のデータの相関係数であるという定義に従えば、その結果を解釈することは比較的容易である。しかし一方で、前節でのメタボロミクス研究における主成分分析の適用を調査した一連の論文においては、図から判断して、因子負荷量の絶対値が 1 に比べかなり小さくなっている。つまり、因子負荷量が先述の定義である主成分スコアと各物質のデータの相関係数とは異なっていることがわかる。この問題について、次に説明する。

まず相関係数を元にした因子負荷量の定義を用いれば、因子負荷量はいくつかの式変形の後、次のように書ける [94]。

$$\mathrm{corr}(\mathbf{z_m}, \mathbf{x_p}) = \frac{\sqrt{\lambda_m} w_{m,p}}{\sigma_{x_p}} \qquad \text{式 (1)}$$

ここで $corr(z_m,x_p)$ は第 m 主成分スコア z_m と物質 p のデータ x_p の相関係数であり、先述の因子負荷量の定義そのものである。λ_m は第 m 主成分の固有値で、主成分スコアの分散である。全主成分に対する λ の和に対する λ_m の割合を寄与率と呼ぶ。$w_{m,p}$ は第 m 主成分の固有ベクトルの物質 p に対応する値、σ_{xp} は物質 p のデータ x_p の標準偏差である。

上式から、次のことが明らかとなる。まず、データを平均 0 分散 1 とする autoscaling を行った場合、上式の分母 σ_{xp} は 1 となることから、因子負荷量と主成分分析の固有値問題から計算される固有ベクトルは比例する。前節で挙げた一連の論文中では、この固有ベクトルが因子負荷量として用いられている。主成分分析においては、固有ベクトルはその長さを示すノルムが 1 である制約条件下で計算されていることから、一般的に式 (1) の定義よりも値として小さくなる。式 (1) の因子負荷量の定義に従うならば、この計算された固有ベクトルに、固有値の平方根である主成分スコアの分散を掛ける必要がある。

一方、autoscaling が行われなかった場合も、先程と同様に、固有ベクトルそのものが用いられているようである。この場合、固有ベクトルは主成分スコアと各物質の相関係数とは比例せず、その共分散と比例する。

$$\mathrm{cov}(\mathbf{z_m}, \mathbf{x_p}) = \lambda_m w_{m,p} \qquad \text{式 (2)}$$

先述の定義に合わせるならば、因子負荷量を求めるためには固有ベクトルに固有値の平方根を掛け、さらに物質 p の標準偏差 σ_{xp} で割る必要がある。

最後に、因子負荷量の言葉の定義についてもう一度整理する。先述の書籍 [89 90 91 92 93] では、因子負荷量は主成分スコアと各物質のデータの相関係数であると定義し、その定義通りの " 因子負荷量 " とは別に、主成分分析の固有値問題の解である固有ベクトルを、" 主成分の係数 "、" 重み " もしくは " 重みベクトル "、" 固有ベクトル " という表現で用いている。 また , 論文等で比較的よく引用される主成分分析の書籍 [50] では , "Some authors

distinguish between the terms ‘loadings’ and ‘coefficients’, depending on the normalization constraint used, …” とある。” 因子負荷量 ” と ” 主成分係数 ” の言葉の違いが , 標準化つまり autoscaling に依存することを断った上で、理論的説明の側面が強い本書では、それらを同一に扱うと説明している。

以上のように、因子負荷量は応用上有用な指標である一方で、その言葉の定義が曖昧であり、その理由について本節で説明した。実際の計算とその結果の解釈を行うに際しては、ソフトウェアのマニュアルなどから、因子負荷量の定義をよく確認しておく必要があるだろう。

4.3 因子負荷量の統計的仮説検定

前節では、データの前処理として autoscaling を行った際に、因子負荷量が主成分スコアと各代謝物レベルの相関係数に相当することを説明した。autoscaling を行わずに主成分分析を行うと、主成分分析、特に因子負荷量が相対値の大きさに影響され、極端に少数の代謝物のみが生物学的な解釈の対象となることから、偏った結論に辿り着く可能性がある。一方、autoscaling を行った場合は、全ての代謝物を平均 0, 分散 1 となるように変数変換を行っていることから、相対値の大きさの影響を受けることはなく、以下で説明するように、適切な数の代謝物を統計的仮説検定を用いて選択出来るという意味において偏りがなく、生物学的な解釈を行う上ではこの方が適切であると考えられる。よって以降では、autoscaling による前処理を前提とし、因子負荷量の仮説検定の方法を用いた有意な代謝物群の選択の方法について説明する。スケーリングの検討に関しては、文献が詳しい[96]。

相関係数の統計的仮説検定は、相関係数 r を用いた統計量

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \text{式 (3)}$$

が、自由度 n-2 の t 分布に従うことを利用する。ここで、n はサンプル数である。実際の計算は Microsoft Excel 等で簡単に計算することができる。実際の手順は、まず主成分スコアプロットから興味あるパターンを示す主成分軸を見つけ、その主成分に対応する因子負荷量を計算する。因子負荷量は、主成分スコアと各代謝物データの相関係数から計算するか、統計解析ソフトウェアから固有値と固有ベクトルが得られている場合には、式（1）に代入することで、相関係数に変換する。得られた相関係数の値を式（3）に代入して t 統計量を計算する。この t 統計量から、例えば Microsoft Excel を用いる場合には関数 ”tdist” を利用して、p-value を計算することが出来る。最後に、多重検定を考慮した p-value の補正や q-value を計算し、これを基準として、注目すべき代謝物群を選択する。次節では、この多重検定の問題について説明する。

4.4　多重検定

メタボロミクスに限らず、オミクスデータの統計的仮説検定は多くの項目について検定を行う多重検定となることから、p-value の補正[97]もしくは False discovery rate (FDR) を基準とした q-value[9899]を用いて有意な物質群を選択する。本節では、メタボロームデータの多重検定についてシミュレーションを交えて説明する。

多重検定の問題を説明するために、次のようなシミュレーションを設定する。2 群で各群のサンプル数がそれぞれ 10、代謝物質数は 500 のデータセットを用意する。このデータセットは、全ての代謝物質で群間の平均に差が無い、つまり全て独立に帰無仮説に従うとする。このデータセットにウェルチの t 検定を適用した結果、5% 有意水準以下で有意となる代謝物質は 20 得られた。つまり全く差が無い代謝物質しか含まれないはずのデータセットに、20 もの統計的に有意な代謝物質が得られることになる。

この多重検定の問題を避けるために、有意水準を 5% より引き下げる、もしくは p 値の補正を行うことがオミクスの統計的仮説検定の適用ではよく行われている。その方法の一つとして、ボンフェローニの補正[97]が用いられる。ボンフェローニの補正では、有意水準を繰り返した検定の数、つまり代謝物質数で割ることにより補正する。この例で言えば、有意水準を 0.05/500=0.0001 とすることに相当する。ボンフェローニの補正を行った場合、有意となる代謝物質数は 0 個となる。一方で、q-value もしくは FDR が用いられることも多い[9899]。q-value を基準とした方法は、ボンフェローニの補正の p-value よりも検出力が高いことが知られている[100]。 q-value は、R の q-value ライブラリ[101]で、また FDR は同じ R の q-value ライブラリから帰無仮説の数の割合であるπ_0の推定量が計算出来るので、有意水準を決めれば、例えば文献[101]にある FDR の推定量の式に代入することで簡単に計算できる。

また多重検定の問題を簡単に確認するためには、p 値のヒストグラムを描けばよい。真に差がある代謝物質が含まれないデータセット中の p 値の分布は、図 4.2(左) のようになりそ

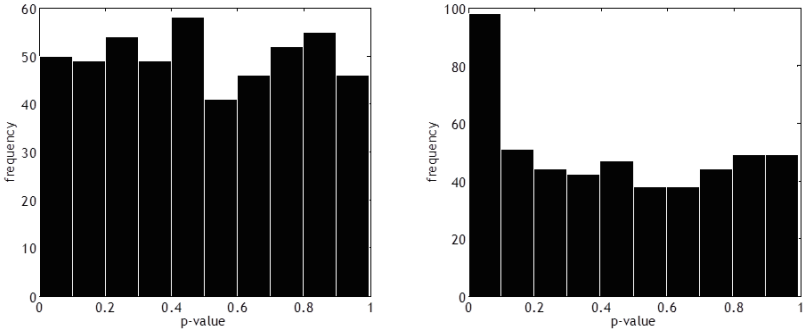

図 4.2　p-value のヒストグラムの例

の分布は一様分布になることが知られている。一方で、差のある代謝物質がいくつか含まれたデータセットの p 値は図 4.2(右) のようなヒストグラムになる。

実際に、全代謝物に対して p-value を計算し、図 4.2(左) のようなヒストグラムが得られた場合には、統計的に有意な物質が得られていないか、非常に少数であることになる。一方で、図 4.2(右) のようなヒストグラムが得られた場合には、様々な物質で表現型の違いが現れており、さらなる生物学的な考察を行う価値があると言えるだろう。

4.5 因子負荷量と metabolite set enrichment analysis

3 節では因子負荷量の統計的仮説検定の方法、4 節では多重検定の問題について説明した。次に、因子負荷量と統計的仮説検定を基準として得られた物質群から、生物学的な解釈を行いたい。生物学者によるメタボロームデータの生物学的解釈を見ていると、例えば " 解糖系が活性化されている " や " アミノ酸類が上昇している " といった表現が用いられている。因子負荷量から得られた代謝物リストを、これらの表現に変換するための統計的な解析が次に行う手順となる。そのための統計的方法として、metabolite set enrichment analysis(MSEA) がある。

現状において、MSEA の基礎となっている統計的方法は、遺伝子発現データで一般的に広く用いられている gene set enrichment analysis(GSEA)[102] と共通である。GSEA の方法として最も有名な方法の一つとして、スブラマニアンらによる方法 [103] が挙げられる。この方法では、遺伝子発現データと群情報などの外部変数との相関係数を基準として降順に並べ、GO(Gene Ontology) を基準とした遺伝子セットの情報を利用して独自のスコアを計算し、p-value もしくは q-value を計算している。もう一つの方法として、over-representation analysis(ORA)[104] がある。ORA は、スブラマニアンらの方法と同様に、GO に基づいた遺伝子セットと統計的仮説検定等によって選択した遺伝子群についての 2 × 2 のクロス集計表から、特定の遺伝子セットと有意な遺伝子群との関連を調べる方法である。

MSEA の手法を、主成分分析の因子負荷量に適用する手順は単純である。スブラマニアンの方法では、各代謝物データと主成分スコアの相関を用いればよく、因子負荷量の値そのものを用いればよい。ORA については、因子負荷量の仮説検定の結果から代謝物質群を選択し、2 × 2 クロス集計表の検定を行う。代謝物質セットの分類は、我々は KEGG[105] から取得し改変したものを利用している。

MSEA の実際の計算としては、我々はインハウスで実装したものを利用しているが、オンラインでフリーで利用できるものとして、MetaboAnalyst[106] や MBRole[107] などがある。 また Ingeuity pathway analysis(IPA) や Metacore 等の商用ソフトウェアにも実装されている。実際に実装されている手法としては、ORA が多いようである。これは、スブラマニアンらの方法はサンプル数が必要であるのに対して、ORA では少数のサンプル数でも計算が可能

であることから、好んで使われていると思われる。

主成分分析における因子負荷量の統計的仮説検定と MSEA の利用により、各代謝物セットごとに統計的仮説検定が行えることを説明した。これにより、本節の初めに述べた " 解糖系が活性化されている " や " アミノ酸類が上昇している " といった表現が、統計的に行えることになる。実際に MSEA の適用例はまだほとんど無く、今後 MSEA を用いた研究が様々報告されていくだろう。

4.6 Partial least squares を用いたメタボロームデータ解析

ここまでは、主成分分析を用いたメタボロームデータ解析の方法について述べてきた。しかしながら、主成分スコアに興味あるクラスターを発見できなかった場合はどうすれば良いだろうか。本節では、メタボロームデータそのものだけでなく、群情報や群の順序の情報、時系列情報が与えられている状況下での、この問題に対する一つの解決策を述べる。

まず最も単純な例として、データに加えて、群情報が与えられている状況を考える。主成分分析と、PLS[51] の結果を図 4.3 に示す。用いたデータは 30 サンプル 500 変数で、そのうち大部分の 480 変数がランダムな変数、残りの 20 変数は群間に差がある変数の合成データを作成し、autoscaling を行った後、主成分分析と PLS の計算をそれぞれ行った。結果より、主成分分析では興味あるクラスターは得られなかったが、PLS では群情報に沿ったクラスターが得られている。これより、PLS の第一軸は群情報を反映した軸である、と解釈できる。

一方、主成分スコアのプロットで、興味あるクラスターが得られなかったというのもまた 1 つの情報である。この場合、第一主成分の寄与率は 4.91% で、第二主成分の寄与率は 4.90% であった。第二主成分までの累積寄与率は約 10% であり、この第二主成分までの累積寄与率が低いという事実は、データに含まれる代謝物質のパターンが非常に多様であることを示唆している。ここで用いたデータセットは、多くがランダム変数であることを考えて

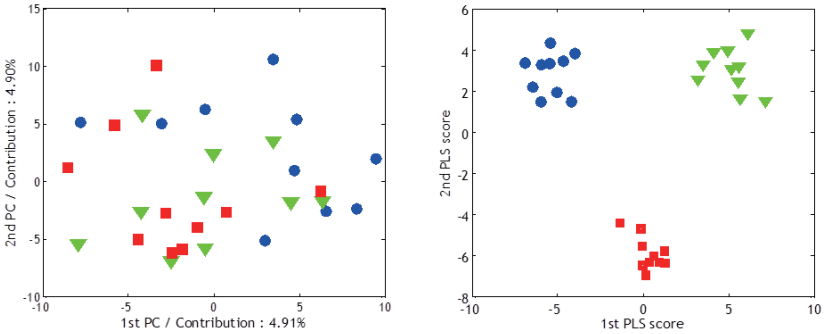

図 4.3　合成データに対する主成分スコア (左), PLS のスコア (右)

も、当然の結果であると言える。また群間に差のあるパターンが第一もしくは第二主成分に現れなかったことから、群間差を示す代謝物質が、少数であることもこの結果から想像がつくであろう。

次に群情報に加えて、群の順序の情報が与えられている状況を考える。例えば疾患のステージとして、{健常, 早期, 後期}や、官能評価での{おいしい, 普通, まずい}といった情報が与えられているとする。合成データを用いた計算結果を図 4.4 に示す。

例えば、●は健常、▼は早期、■は後期と考える。図 4.4(左)は PLS の結果を示しており、青→緑→赤の順序を示すパターンは、スコアプロットには現れていない。この場合に、群の順序を考慮した PLS-ROG[108] を用いると、図 4.4(右)の結果が得られ、第一軸が、疾患のステージに関連する軸であると解釈できる。さらに因子負荷量から疾患のステージと相関する物質を特定することも出来る。

実際に、この方法を用いて、CE-MS を用いたワインのメタボローム解析において、年代と相関するスコアを見つけ出すことに成功している[109]。図 4.5 に結果を示す。●は 2000、＋は 2001、▼は 2004、×は 2005、＊は 2007、■は 2008 年のワインのサンプルであり、左は PLS、右は PLS-ROG の結果である。主成分分析の結果は、PLS の結果と類似していたことから、ここでは省略する。

主成分分析または PLS では、第一軸で{2000, 2001 年}と{2004, 2005, 2007, 2008 年}の 2 つのクラスターに分かれている。この違いは、ワインの醗酵法の違いに依存していることが、その後の考察で明らかとなった[109]。また、PLS-ROG の結果(図 4.5(右))は、第一軸が年代を表していることから、因子負荷量から年代と相関する物質を特定し、MSEA からさらなる生物学的な解釈を行うことが可能となる。ただし、PLS の因子負荷量は主成分分析とは量的な意味が異なるので、注意が必要である[110]。

その他、醗酵プロセス等において、サンプルに時系列の情報が付加情報として与えられている場合に適した主成分分析も提案されている[111]。このように、従来の多変量解析を用

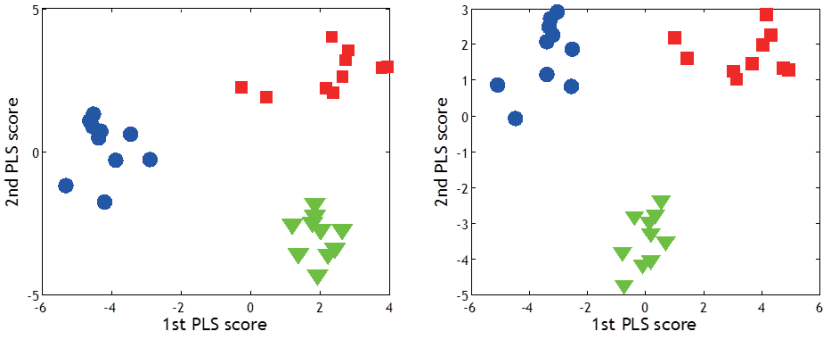

図 4.4　合成データに対する主成分スコア(左), PLS-ROG のスコア(右)

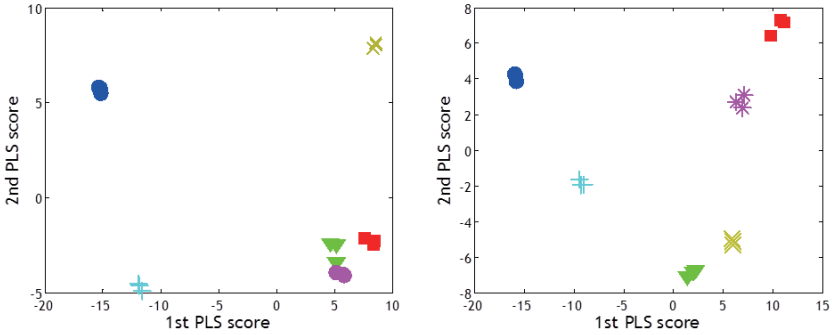

図 4.5　ワインのデータに対する PLS のスコア (左), PLS-ROG のスコア (右)

いて興味のあるクラスターが見られなかった場合でも、本節で紹介したような手法を用いれば、新たな結果が得られることが期待される。

4.7　予測モデル構築

メタボロミクスでは、多変量解析を予測モデルの構築を目的として適用する研究も行われている。その中の例の一つとして、回帰分析を用いた予測モデル構築の研究例 [112][113] を紹介する。

品評会において 1 位から 53 位までランク付けされた 53 の緑茶がある。この緑茶の葉を GC-MS を用いてメタボローム解析を行い、メタボロームデータのみから緑茶の味の評価値を予測するモデルを PLS 回帰または正準相関分析 *[113] を用いて構築した。これにより、新たな緑茶の味の評価を、メタボロームデータから予測することが出来る。この他にも、ロジスティック回帰を用いた唾液サンプルのメタボロームデータからのがん診断モデルの構築 [114] など、様々な研究が行われている。

回帰の予測モデル構築では、ケモメトリクス **[115] 分野で良く用いられている PLS 回帰 [116] が用いられることが多い。これは前節で説明した PLS[51] とは理論的に異なっている。PLS[51] と、PLS 回帰 [116] もしくは PLS-DA の一番の違いは、後者はそのスコアが直交であるとする制約条件が課せられている点にある。特に回帰において、説明変数が直交であることは、線形重回帰分析での予測性能の低下、もしくは逆行列の計算において、数値計算の不安定を引き起こす、多重共線性の問題を回避することが出来る。その他、Orthogonal Projections to

*　説明変数であるメタボロームデータの合成変数と、目的変数の合成変数の相関係数最大化を基準とした教師あり次元削減法。p(物質数)>>n(サンプル数) のオミクスデータに適用する際は、l_1 もしくは l_2 正則化項を用いた正則化正準相関分析が用いられる。

**　化学分析データの中でも特に近赤外スペクトルデータに、PLS 回帰を初めとする多変量解析の方法を用いて解析することに積極的な研究分野。

Latent Structures(OPLS)[117] や正準相関分析 [113] も用いられている。これらの主な利点は、モデルに用いるスコアの数が少なく、PLS と同程度の予測精度のモデルを構築でき、これにより因子負荷量の解釈がしやすい点にある。

4.8 おわりに

従来の多変量解析特に主成分分析の利用法では、主成分スコアからサンプルがどのようなパターンを示すのか、また外れ値の存在の確認といった程度にしか使われておらず、様々なメタボロミクスの論文を読んでいる中で、生物学的な解釈までは大きなギャップがあると感じていた。その理由の一つとして、因子負荷量の量的な意味と、その利用方法が、メタボロミクス研究に十分浸透していないのではないかと考えている。

そこで本章では、特に因子負荷量について詳細に説明し、多変量解析の結果から、生物学的な解釈までの一連の流れを統計解析を用いて行う方法を紹介した。実際にメタボロームデータをお持ちの研究者の方々やこれからメタボロミクスを行う予定の研究者の方々には、本章を参考に統計解析を行っていただければ幸いである。