HDLSS 環境における先進的ケモメトリクス：スペクトル前処理から数理統計学まで

Yugo Nakayama

2026 年 2 月 10 日

概要

本サーベイ論文は、高次元小標本 (HDLSS) 環境に焦点を当てたケモメトリクスの包括的な概要を提示する。スペクトル前処理の基礎から、古典的な多変量解析手法 (PCA, PLS)、そして先進的なスパースモデリングアプローチまでを体系的に網羅する。さらに、強スパイク固有値 (SSE) モデルやノイズ削減法といった厳密な高次元統計理論を統合し、化学データ解析に対する深い数学的理解を提供する。NIR 分光法やマイクロアレイ解析におけるケーススタディと共に、Python と R による実践的な実装も示し、理論と実践の架け橋となることを目指す。

目次

1 ケモメトリクスの基礎 4
1.1 定義と範囲：化学とデータの架け橋 4
1.2 化学データ特有の「癖」: なぜ専門の手法が必要か 4

2 歴史的変遷と最近の動向 5
2.1 ケモメトリクスの進化と歴史的文献 (1970 年代–現在) 5
2.2 高次元統計学 (HDLSS) の発展 6

3 スペクトル前処理と補正 7
3.1 なぜ前処理で「邪魔者」を消さないといけないのか？ 7
3.2 「化学的シグナル」とは何か？ 7
3.3 代表的な前処理手法の直感的理解 7
3.4 SNV と Savitzky-Golay 法の違いと使い分け 9
3.5 前処理なしの失敗例 9

4 主成分分析 (PCA) ー ケモ的な意味付け 9
4.1 PCA で見えているもの 9
4.2 ガソリン PCA のイメージ（図 2 の読み方） 10
4.3 ケモ的 PCA の使いどころ 11
4.4 HDLSS での PCA の位置づけ 12

5 主成分回帰 (PCR) と部分的最小二乗法 (PLS) 12
5.1 主成分回帰 (PCR) 13
5.2 部分的最小二乗法 (PLS) 13
5.3 PCR と PLS のケモメトリクス的比較 14

6 PLS の発展的応用 14
6.1 PLS 判別分析 (PLS-DA) 14
6.2 スパース PLS (Sparse PLS) と変数選択 15

7 QSAR/QSPR と分子記述子 15
7.1 分子記述子 15

8 定量的構造活性相関 (QSAR/QSPR) 15
8.1 分子記述子：構造をどうやって「数字」にするか 15
8.2 PLS-QSAR モデル 17
8.3 モデル検証統計量：R^2 と Q^2 の意味 17

9 HDLSS 環境における統計理論 18
9.1 この章で扱うこと（導入） 18
9.2 強スパイク固有値 (SSE) モデル 18
9.3 ノイズ削減 (NR) 法 19

10 実装とソフトウェア環境 19
10.1 再現可能なコード資産 20
10.2 実践例：ガソリン分析 20

11 ケーススタディ 21
11.1 ケース 1: ビールの NIR 分析 (iPLS) 21
11.2 ケース 2: ガソリングレード分類 (PLS-DA) 21

12 今後の展望と課題：研究ロードマップ 22
12.1 テンソル分解（マルチウェイ解析） 23
12.2 ケモメトリクスにおける深層学習 24
12.3 因果推論 25

付録 A ベンチマークデータセット 26
A.1 代表的な分光データセット 26
A.2 その他の有名なデータセット 27
A.3 パブリックデータの比較と活用指針 27
A.4 データセットの入手方法 27

付録 B 標準的な解析ワークフロー 28
B.1 フェーズ 1: データの構造を知る（探索的データ解析） 28
B.2 フェーズ 2: サンプルの分割（トレーニングとテスト） 28
B.3 フェーズ 3: モデルの構築（キャリブレーション） 28
B.4 フェーズ 4: モデルの検証（バリデーション） 29
B.5 フェーズ 5: 解釈と応用 29

付録 C ケモメトリクスのための数学的準備 29
C.1 行列とベクトルの記法（もう一歩ていねいに） 29
C.2 行列形式での基本統計（直感補強） 30
C.3 固有分解と PCA とのつながり 31

C.4 線形回帰 (OLS) とその限界の明示 31

付録 D ケモメトリクス手法の理論的定式化 31
D.1 主成分分析 (PCA) 32
D.2 部分的最小二乗法 (PLS) 32
D.3 主成分回帰 (PCR) 33

1　ケモメトリクスの基礎

ケモメトリクス（計量化学：Chemometrics）は、一言でいえば「化学データのためのデータサイエンス」である。化学実験や分析装置から得られる膨大なデータから、数学、統計学、コンピュータサイエンスの力を借りて、目には見えない有用な情報を効率よく引き出す学際的な学問分野である。

1.1　定義と範囲：化学とデータの架け橋

ケモメトリクスは、単なる「計算」ではない。実験の計画段階から、得られた結果の解釈、さらには未知の現象の予測まで、化学研究のあらゆるプロセスを最適化する。主な役割は以下の 4 点に集約される。

- **最適な実験の設計（実験計画法）**：「当てずっぽう」に条件を変えるのではなく、最小の実験回数で最大の情報を得るための条件設定（温度、濃度、反応時間など）を数学的に導き出す。
- **情報の最大化（信号処理）**：分析装置から出てくる生のデータには、微弱な化学シグナルと多くのノイズが混ざっている。これらから「意味のある情報」だけを鮮明に取り出す技術である。
- **未知の特性を当てる（キャリブレーションと回帰）**：「光の吸収パターン（スペクトル）」から、その物質の「濃度」や「美味しさ（糖度など）」を瞬時に予測する計算式（モデル）を作る。一度モデルができれば、化学分析の手間を省き、リアルタイムでの測定が可能になる。
- **グループを見分ける（パターン認識）**：「この産地のワインはどれか？」「この薬は本物か偽物か？」といった問題を、多成分のデータから自動的に判別する。

1.2　化学データ特有の「癖」：なぜ専門の手法が必要か

一般的な統計学（例えば身長と体重の関係を調べるようなもの）と違い、化学データ、特に光を使って測定する「スペクトルデータ」には、古典的な統計手法では太刀打ちできない特有の性質がある。

- **高次元小標本 (HDLSS) 問題**: 例えば、1 枚のスペクトルを撮ると、数千箇所の波長のデータ（変数 d）が得られる。一方で、準備できるサンプルの数（サンプル数 n）は、コストや時間の制約から数十個程度であることが珍しくない。このように「情報の項目数が、データの件数より圧倒的に多い $(d \gg n)$」状態は、従来の統計学では計算が破綻する原因となる [10]。
- **多重共線性（マルチコリニアリティ）**：スペクトルのグラフを思い浮かべると、隣り合った波長は似たような動きをする。統計学的に見ると「変数同士が非常に似通っている（強い相関がある）」状態である。これにより、普通の回帰分析（最小二乗法）を行うと、計算結果が極端に不安定になり、信頼できないモデルになってしまう [17]。
- **ノイズとベースラインの変動**: 測定時の室温の変化、試料の粒の大きさ、装置のわずかな調子の違いによって、グラフ全体が上下にズレたり（ベースラインシフト）、ザラザラした雑音（ノイズ）が乗ったりする。これらは化学的な成分とは無関係な「邪魔者」であり、これらを数学的に取り除く「前処理」が極めて重要になる [3]。

このように、化学データは非常に「癖が強い」ため、それらを逆手に取って効率よく解析する **PLS（部分的最小二乗法）** や **PCA（主成分分析）** といった、ケモメトリクス独自の強力な武器が必要とされるのである。

2　歴史的変遷と最近の動向

本稿では、ケモメトリクスと高次元統計学の融合領域について探求する。両分野の現在の収束を理解するためには、それぞれの歴史的軌跡を理解することが不可欠である。

2.1　ケモメトリクスの進化と歴史的文献 (1970 年代–現在)

ケモメトリクスの歴史は、分析装置の高度化によって生じた「情報の洪水」を、いかにして意味のある化学的知見へと凝縮するかという挑戦の歴史である。

2.1.1　創成期：多変量解析の夜明け (1970 年代)

この時代は、単変量から多変量へ、そして「データの幾何学的解釈」への移行が起きた時期である。

- **多重共線性への数学的回答**: 初期の NIR や GC-MS データでは、隣接波長間が高い相関を持つため、設計行列 $\mathbf{X}$ がランク落ち（Rank-deficient）に近い状態になる。OLS（最小二乗法）では $(\mathbf{X}^T\mathbf{X})^{-1}$ が発散し、回帰係数の推定値が不安定になる。**Herman Wold** が計量経済学のために開発した **NIPALS** アルゴリズム [15] は、行列の固有値分解を介さず、反復的な射影によって主成分や潜在変数を抽出することで、この特異性の問題を実用的に解決した。
- **ケモメトリクスの定義**: **Svante Wold** は 1972 年に “Chemometrics” という用語を初めて公式に使用し [16]、**Bruce Kowalski** はパターン認識の概念を化学に持ち込んだ [11]。これは、単なる定量だけでなく、未知サンプルの「種別」や「起源」を多変量空間上の距離で定義する学問としての基礎を築いた。
- **PLS の体系化**: Wold らは、化学における多重共線性問題を PLS がいかに解決するかを論理的に体系化した [17]。これは実質的な「PLS の教科書」として機能した。

2.1.2　拡大と標準化：NIR の産業応用と前処理技術 (1980–1990 年代)

1980 年代に入ると、農業や食品分野における非破壊検査として NIR が普及した。この時代は、ラボレベルの数理を「現場で使える技術」へと昇華させるための標準化が進んだ。

- **散乱補正の理論化 (SNV と MSC)**: 粉体試料を測定する際、光の経路差によりスペクトル全体がシフトする「乗法的効果」が問題となった。**Geladi ら (1985)** は平均スペクトルを基準とした回帰による **MSC (乗法散乱補正)** を提案し [9]、**Barnes ら (1989)** は個別のスペクトル統計量に基づく **SNV (標準正規変量)** を提案した [3]。これにより、化学情報（シグナル）と物理的ノイズの分離が可能となった。
- **実務的ガイドライン**: **Geladi と Kowalski (1986)** は実務家向けに PLS を解説したチュートリアル論文を発表し [8]、交差検証や RMSEP によるモデル評価の文化を定着させた。

2.1.3　現代：高度な情報の分離と高次データ解析 (2000 年代以降)

オミクス技術の台頭により、データは「多次元」かつ「マルチウェイ」なものへと進化した。

- **情報の直交分離 (O-PLS)**: **Trygg と Wold (2002)** が発表した **O-PLS** は、予測に寄与する変動と、それ以外の直交（Orthogonal）成分を数学的に分離し、モデルの解釈性を劇的に向上させました [14]。
- **マルチウェイ解析**: 三次元励起蛍光スペクトル (EEM) などのテンソルデータに対し、**Rasmus Bro (1997)** は **PARAFAC** モデルを化学分野に定着させた [4]。これにより、数学的なクロマトグラフィー（成分抽出）が可能となった。

- **高次元統計理論 (HDLSS)**: **Hall ら (2005)** は高次元データの幾何学的性質（測度の集中）を証明し [10]、**青嶋と矢田 (2018)** は標本固有値のバイアスを補正する厳密な理論的枠組み (SSE モデル) を構築した [2]。

表 1　歴史を象徴する主要文献一覧

時代	貢献内容	主要論文（筆頭著者, 年）
1970s	パターン認識の導入	Kowalski and Bender [11]
1980s	PLS の実用チュートリアル	Geladi and Kowalski [8]
1985	散乱補正 (MSC) の提案	Geladi et al. [9]
1989	散乱補正 (SNV) の提案	Barnes et al. [3]
1997	テンソル解析 (PARAFAC)	Bro [4]
2002	情報の直交分離 (O-PLS)	Trygg and Wold [14]
2005	HDLSS の幾何学的性質	Hall et al. [10]
2018	SSE モデルと不偏推論	Aoshima and Yata [2]

2.2　高次元統計学 (HDLSS) の発展

ケモメトリクスと並行して、数理統計学においても「次元の呪い」に対処するための革命が起きていた [6]。

2.2.1　HDLSS データの幾何学的表現 (2005)

画期的なブレークスルーは、**Hall, Marron, Neeman (2005)** によってもたらされた。彼らは、n を固定したまま $d \to \infty$ としたときのデータの幾何学的性質を数学的に特徴づけた [10]。彼らは、穏やかな条件の下で、データ点は正単体（regular simplex）を形成する傾向があり、平均までの距離が決定的になる（「測度の集中」）ことを証明した。これにより、ノイズ構造に応じて距離ベースの分類器（k-NN など）が驚くほど機能したり、あるいは壊滅的に失敗したりする理由の厳密な基礎が与えられた [1]。

2.2.2　スパースモデリングと変数選択

解釈性と一致性を達成するため、統計的学習はスパース性に目を向けた：

- **Lasso (1996)**: Tibshirani は L_1 正則化を導入し、回帰と変数選択の同時実行を可能にした [13]。
- **オラクル性 (2001)**: Fan と Li は、Lasso に内在するバイアスを補正し、「オラクル」特性（真のモデルを知っているかのように振る舞うこと）を保証するために、SCAD のような非凸ペナルティを提案した [7]。

これらの発展はケモメトリクスに直接的な影響を与え、**Sparse PLS** へとつながった [5]。

2.2.3　最近の進展：PCA の一致性と SSE モデル (2010 年代)

より最近では、**Aoshima と Yata** が、正規性やスパース性を仮定せずに HDLSS 推論を行うための包括的な枠組みを開発した。彼らの **Strong Spiked Eigenvalue (SSE)** モデルと **Noise Reduction (NR)** 推定法 [18, 2] は、超高次元における固有値と固有ベクトルの不偏推定量を提供し、Marchenko-Pastur 則によって記述される標本 PCA 固有値の体系的なバイアス（不一致性）に対処している。

3　スペクトル前処理と補正

本節では、「前処理」の必要性と、「化学的シグナル」の正体について、直感的なイメージを用いて解説する。

3.1　なぜ前処理で「邪魔者」を消さないといけないのか？

結論から言えば、**「本来知りたい情報（化学的な中身）」が、「見た目の違い（物理的な状態）」に完全に埋もれてしまうから**である。

日常的な例として、以下の 2 つの液体を比較することを考える：

1. **薄い**カルピスを、**とても太い**コップに入れたもの
2. **濃い**カルピスを、**とても細い**ストローに入れたもの

これらに真横から光を当てて測定すると、「光の通り道」が長い 1 番の方が、光がたくさん吸収されて「濃い」という誤ったデータが出てしまうことがある。これは**コップの太さ（物理的要因）**のせいで、**中身の濃さ（化学的要因）**を正しく測れていない状態である。

3.1.1　前処理をしないと起こる問題

- **「見た目」が「中身」に勝ってしまう**: 光の散乱（試料の粒の大きさの違い）やベースラインのズレ（装置の調子の差）は、しばしば化学的なシグナルよりも**数十倍も大きなデータ上の変化**として現れる。前処理をしないと、モデルは「粒が粗いか細かいか」だけを見て判断し、肝心の「成分が何か」を無視してしまう。
- **偽の相関（Spurious Correlation）**: 例えば、「高級なワインの瓶がたまたま少しだけ厚かった」とする。前処理をしないと、モデルは「ガラスの厚み」を「高級さの証」だと勘違いして学習してしまう。これでは、中身が偽物でも瓶さえ厚ければ「本物」と判定する、実用性のないモデルになってしまう。

3.2　「化学的シグナル」とは何か？

化学的シグナルとは、**「その分子だけが持つ固有の反応（指紋）」**のことである。具体的には、以下のような「分子の動き」を指す：

- **振動**: 分子は常にバネのように伸び縮み（伸縮振動）したり、曲がったり（変角振動）している。
- **吸収**: 例えば、アルコールに含まれる -OH 基は、特定のエネルギーの光を吸収して激しく動く。

このシグナルを捕まえることで、分子の種類を特定（定性分析）したり、その強さから分子の数（濃度）を計算（定量分析）したりすることが可能になる。

3.3　代表的な前処理手法の直感的理解

前処理の目的は、データの「見た目のデコボコ（物理的なノイズ）」を整えて、「本当の中身（化学的な成分）」を見えやすくすることである。

3.3.1　1. 散乱補正 (SNV / MSC)：コップの太さを揃える

光の散乱の影響を取り除く手法である [3, 9]。

- **イメージ**: 「カルピスの濃さテスト」

- **状況**: 入れ物が「細いストロー」だったり「太いジョッキ」だったりバラバラだと、正しく濃さが測れない。
- **処置**: 全てのデータの平均値を基準にしたり、標準偏差で割ったりして、データの「スケール」を一定に揃える。
- **結果**: 器の形に関係なく、中身の濃さだけで比較できるようになる。

3.3.2　2. 微分 (Derivative)：重なりを切り分ける

重なり合ったピークを分離し、ベースラインを除去する手法である [12]。

- **イメージ**: 「山並みのスケッチ」
- **状況**: 遠くの山（成分 A）の手前に丘（成分 B）が重なっており、さらに地面全体が坂道（ベースライン）になっている。
- **処置**: グラフの「傾き」に注目する（微分）。平らな坂道は「0」になり、重なっていた急な山の輪郭だけが浮き上がる。
- **結果**: 混ざり合っていた成分の「境目」がはっきりする。

3.3.3　3. ベースライン補正 (ALS)：背景の「底上げ」を引く

装置の熱などでグラフが浮き上がってしまう現象を直す。

- **イメージ**: 「上げ底のお弁当箱」
- **状況**: おかずの量は同じなのに、容器が上げ底になっていて見た目の量が変わっている。
- **処置**: グラフの「一番底」を通る曲線を推定し、全体から引き算する。
- **結果**: 全てのデータがゼロからスタートし、純粋な成分量だけで比較できる。

3.3.4　4. 中心化 (Centering)：テストの点数を調整する

データの平均を 0 にする基本操作である。

- **イメージ**: 「**5** 教科の成績表」
- **状況**: 数学は 100 点満点、小テストは 10 点満点の場合、そのまま合計すると数学の影響力が強すぎる。
- **処置**: 全員の平均を「0 点」として揃える。
- **結果**: どの変数の変化も平等に扱えるようになる。

表 2　前処理の手法と直感的イメージのまとめ

手法名	解決したい問題	高校生向けイメージ
散乱補正 (SNV/MSC)	サンプルの大きさや入れ物の違い	器のサイズを揃える
微分 (Derivative)	ピークの重なり、ダラダラした坂道	輪郭をクッキリさせる
ベースライン補正	装置の熱などによる底上げ	上げ底を引いて平らにする
中心化 (Centering)	データの基準点のズレ	平均をゼロにして並べる

3.4 SNV と Savitzky-Golay 法の違いと使い分け

実務において最も頻繁に使用される **SNV** と **Savitzky-Golay (SG) 法** の違いを整理する。一言でいえば、**SNV は「縦のズレ」を直し、SG 法は「横の変化（形）」を鮮明にする**手法である。

- **SNV (Standard Normal Variate)**:
  - **役割**: 全体のスケール（明るさ）を揃える。
  - **対象**: 試料の粒径や光路長による「スペクトル全体の浮き沈み」。
  - **イメージ**: 日向で撮った写真と日陰の写真を同じ明るさに補正する。
- **Savitzky-Golay (SG) 法**:
  - **役割**: ノイズを消し、特徴（ピーク）を際立たせる。
  - **対象**: データのザラザラしたノイズや、ピークの重なり。
  - **イメージ**: ぼやけた写真の輪郭をペンでなぞってクッキリさせる。

黄金パターン: 実際の解析では、これらを組み合わせることが多い。まず SNV で全体の「明るさ（スケール）」を揃え、次に SG 法（微分）で細部の「輪郭（ピーク）」を浮き上がらせるという順序が一般的である。

3.5 前処理なしの失敗例

前処理を怠ると、「全く無関係な要因を化学的な正解だと勘違いする」という致命的なミスにつながる。

1. **「容器の厚み」を「成分の濃さ」と勘違いする**: 散乱補正をしなかったため、AI が「成分の光吸収」ではなく「容器の厚みによる光の減衰」を濃度として学習してしまう。
2. **「工場の室温」で「製品の品質」を判定する**: ベースライン変動を放置したため、装置が冷えている朝は「高品質」、熱を持った午後は「低品質」と誤判定する。
3. **「不純物の微かな変化」を見逃す**: 微分処理をしなかったため、巨大なピークの影に隠れた微小な異常（不純物）を見落とし、不良品を出荷してしまう。

4 主成分分析 (PCA) ― ケモ的な意味付け

PCA は、「たくさんの波長や成分を、一目で理解できる少数の“軸”にまとめる道具」である。ケモメトリクスでは、主に次の 2 つの目的で使用される。

- **探索的データ解析**: 試料同士がどのくらい似ているか・違うかをざっくり眺める。
- **化学的解釈**: その違いを生み出している**波長帯や成分**を特定する。

本節では、数式よりも「PCA を使うと何が見えるのか」を中心に解説する。厳密な数学的定式化（固有値分解や双対性）については **付録 付録 D** を参照されたい。

4.1 PCA で見えているもの

中心化データ $\mathbf{X}_c$ に PCA をかけると、次の 2 種類の情報が得られる。

1. **スコア (Scores)**: $t_{i1}, t_{i2}, \ldots$
   各サンプル i が、主成分 1 軸・2 軸上のどこに位置するかを表す座標。**「試料の個性を、数本の“味の軸”で表した座標（サンプルマップ）」**であり、似た試料は近くに、異なる試料は遠くに配置される。
2. **ローディング (Loadings)**: $p_{j1}, p_{j2}, \ldots$

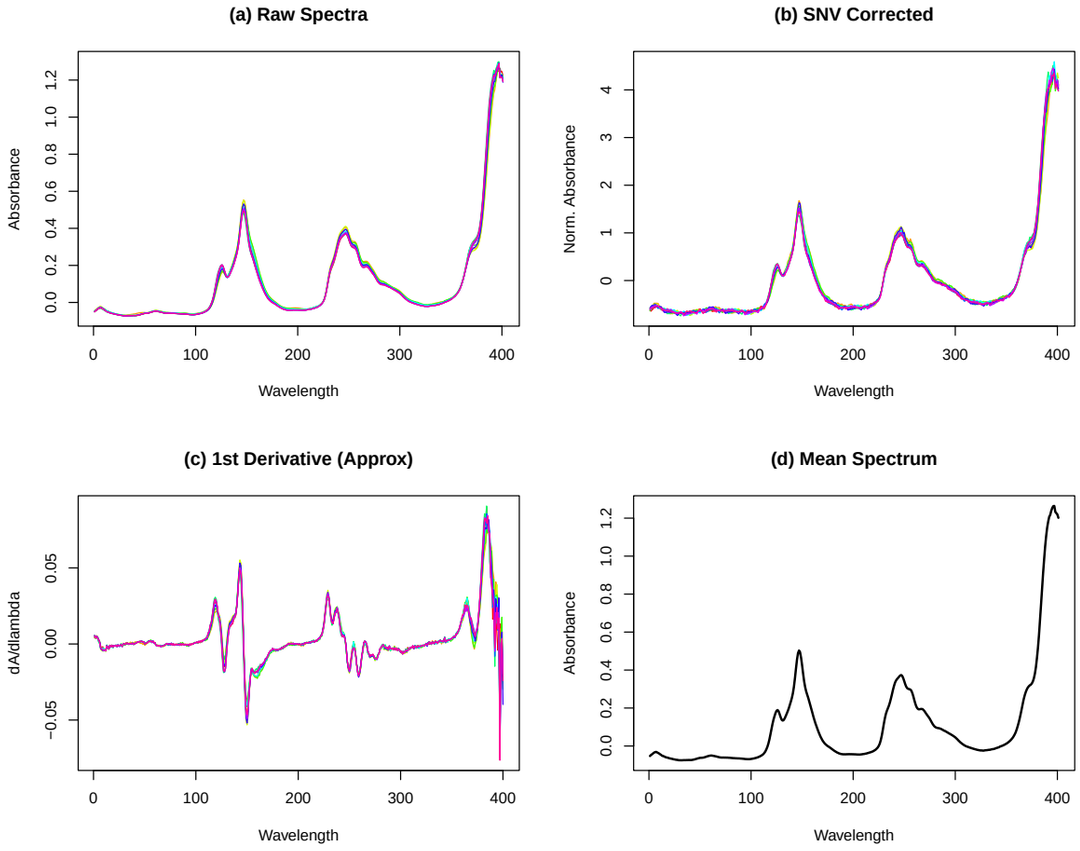

図 1　スペクトル前処理の効果。(a) ベースラインオフセットを含む生スペクトル（「上げ底」の状態）。(b) SNV 補正によりスケールが正規化され、比較可能になる（「器を揃える」）。(c) 一次微分によりピークが強調され、オフセットが除去される（「輪郭を浮き上がらせる」）。

各変数（波長 j）が、主成分 1 軸・2 軸にどれだけ寄与しているかを表す係数。**「その軸を決めている波長・成分（変数マップ）」**であり、どの波長帯が違いを作っているかがわかる。

4.2　ガソリン PCA のイメージ（図 2 の読み方）

図 2 の PCA バイプロットでは、次の 3 つを同時に見ている。

4.2.1　スコア（サンプル側）の解釈

点は各ガソリンサンプルを表す。

- PC1 軸（横軸）上で右側にあるサンプルほど、「ある特定の成分（例えば芳香族）」が多く、左に行くほど少ない傾向がある。
- 似たレシピで作られたガソリン（例：高オクタン・プレミアムガソリン群）は、スコアプロット上でまとまったクラスター（集団）を作る。

化学的には、「PCA スコア図は、401 次元の NIR スペクトル空間を、2〜3 本の“混合割合の軸”に圧縮した地図」と見なせる。

4.2.2　ローディング（波長側）の解釈

矢印は各波長（NIR の測定点）を表す。

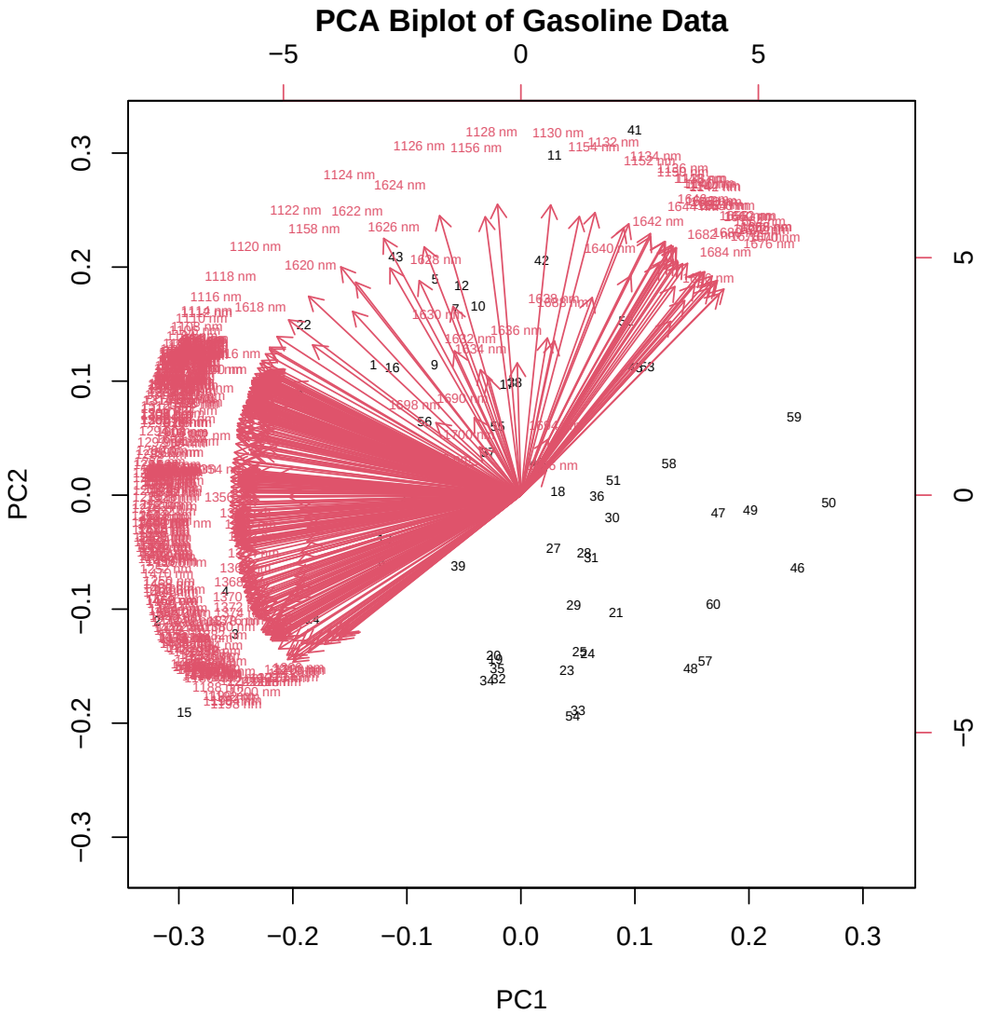

図 2　ガソリンデータの PCA バイプロット。点（スコア）は各ガソリンサンプル、矢印（ローディング）は各波長の寄与を表す。PC1 は主にオクタン価の違い、PC2 は副次的な組成差を反映している。

- **向き**: 原点から遠く、同じ方向を向いている矢印同士は強く相関しており、同じ化学成分や官能基（C-H, O-H など）に由来する可能性が高い。
- **長さ**: PC1 方向に長い矢印は、その波長帯がサンプルの並び（グレード差など）を決める決定的な要因であることを示している。

実務的には、ローディングプロットを見て「どの波長帯が品質差に効いているか」を特定し、その帯域だけを使った簡易モデル（波長選択）を作ったり、ピーク帰属を行ったりするのに使われる。

4.3　ケモ的 PCA の使いどころ

PCA は、現場では次のような問いに答えるために使われる。

- **サンプルのチェック**: 外れ値（測定ミス・異常ロット）はどれか？ 同じ条件で作ったはずなのに、離れているサンプルはないか？
- **隠れたグループの発見**: 産地違い・ロット違い・処理条件違いなどが、自然にクラスターとして現れていないか？

- **重要な波長帯・成分の特定**: どの波長帯がスコアの分離（PC1, PC2）を支配しているか？

4.4　HDLSS での PCA の位置づけ

HDLSS（波長の数 ≫ サンプル数）の NIR データでは、生の 401 次元空間でサンプル間の距離や向きを直感的に理解することは不可能に近い。PCA は、ノイズ的な方向（ランダムな揺らぎ）を捨て、物理・化学的な意味を持つ少数の方向（主成分）だけを取り出すことで、複雑なスペクトル空間を、化学者が読み解ける「**地図**」に変換する技術である。

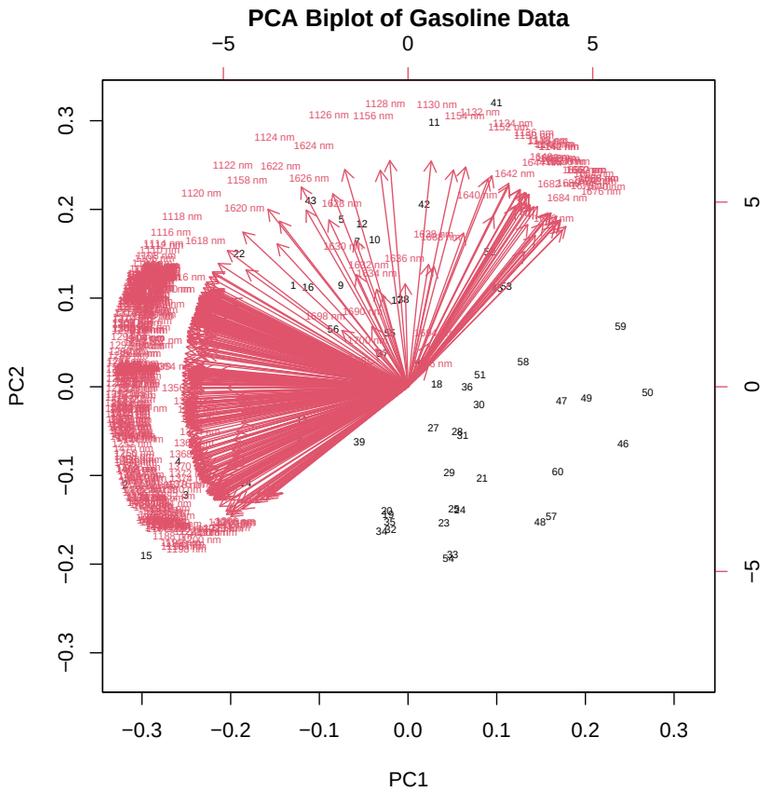

図 3　PCA バイプロット。点はサンプルを表し、矢印は変数（波長）を表す。直観的には、**角度**が相関を表し、同じ方向を向いている矢印は相関している。矢印の方向に位置するサンプルは、その変数の濃度が高いことを示す。このマップにより、化学者はどの化学的特徴が特定のサンプルグループを定義しているかを即座に把握できる。

5　主成分回帰 (PCR) と部分的最小二乗法 (PLS)

HDLSS 環境 ($d \gg n$) では、波長ごとに普通の線形回帰 (OLS) を行おうとすると、以下の理由で破綻する。

- 変数同士が強く相関している（多重共線性）。
- 変数の数がサンプル数より多い（$\mathbf{X}^T\mathbf{X}$ が特異）。

その結果、係数が不安定になり「学習データには合うが、未知データは全く予測できない（過学習）」状態に陥る。PCR と PLS は、**まずスペクトルを少数の“潜在変数”にまとめてから回帰する**ことで、この問題を回避する枠組みである。

ここでは、両者の本質的な違いを以下の通り強調する。

- **PCR**: 「$\mathbf{X}$ の構造」だけを見て軸を作る。
- **PLS**: 「$\mathbf{X}$ と $\mathbf{y}$ の関係」も見ながら軸を作る。

5.1 主成分回帰 (PCR)

5.1.1 手順と数式の意味

PCR は、「PCA でスペクトルを圧縮する（教師なし）」と「圧縮した座標に対して OLS をする（教師あり）」という 2 段階法である。

1. **次元削減**: 中心化された予測子行列 $\mathbf{X}$ に対して PCA を実行し、最初の A 個の主成分を選ぶ。

$$\mathbf{T}_A = \mathbf{X}\mathbf{P}_A \tag{1}$$

   ここで $\mathbf{T}_A$ は主成分スコア（低次元座標）、$\mathbf{P}_A$ はローディング行列である。重要なのは、PCA は「$\mathbf{X}$ のばらつきをよく説明する方向」を選んでいるだけであり、**まだ $\mathbf{y}$ の情報は一切使っていない**点である。

2. **回帰**: 応答 $\mathbf{y}$ を、スコア $\mathbf{T}_A$ に対して OLS で回帰する。

$$\hat{\mathbf{y}} = \mathbf{T}_A \mathbf{b}_{pcr} = \mathbf{T}_A(\mathbf{T}_A^T\mathbf{T}_A)^{-1}\mathbf{T}_A^T\mathbf{y} \tag{2}$$

   主成分は互いに直交しているため、数値的安定性が保証される。

5.1.2 ケモ的な解釈と利点・欠点

- **利点**: スペクトルの「ノイズ的な方向」を自動的に捨て、滑らかな潜在変数に集約できる。$A \ll d$ ならば HDLSS でも安定して計算できる。
- **欠点（本質的な限界）**: PCA の目的は「$\mathbf{X}$ の再現」であり、「$\mathbf{y}$ の予測」ではない。$\mathbf{X}$ の分散が小さい方向（PCA で捨てられる方向）が、実は $\mathbf{y}$ と非常に強く相関している場合、PCR はその重要な情報を**切り捨ててしまう**。

ケモメトリクス的には、「PCR はスペクトルとして“見た目によく効率よく説明できる方向”を優先するが、目的変数にとって役に立つかどうかは二の次」という手法である。

5.2 部分的最小二乗法 (PLS)

PLS は、PCR の「$\mathbf{y}$ を無視して軸を選んでいる」という欠点を補うために、**軸の選び方の段階から $\mathbf{y}$ を意識的に使う**ように設計された手法である。

5.2.1 目的関数の意味

最初の PLS 重みベクトル $\mathbf{w}_1$ は、以下の最適化問題の解として定められる。

$$\max_{\mathbf{w}} \mathrm{cov}(\mathbf{X}\mathbf{w}, \mathbf{y})^2 = \max_{\mathbf{w}} \mathrm{var}(\mathbf{X}\mathbf{w})\mathrm{corr}(\mathbf{X}\mathbf{w}, \mathbf{y})^2 \quad \text{s.t.} \quad ||\mathbf{w}|| = 1 \tag{3}$$

ここで $t = \mathbf{X}\mathbf{w}$ と置くと、t は「スペクトルのある線形結合（潜在変数）」であり、以下の 2 つを同時に大きくする方向を探している。

- $\mathrm{var}(t)$: 潜在変数がどれだけばらついているか（PCA 的な要素）。
- $\mathrm{corr}(t, \mathbf{y})$: 目的変数との相関がどれくらい強いか（回帰的な要素）。

PCA との対比:

- **PCA**: $\max \mathrm{var}(\mathbf{X}\mathbf{w})$ だけを見る。

- **PLS**: var と corr の双方を考慮し、「**よくばらつき、かつ $\mathbf{y}$ とよく一緒に動く方向**」を選ぶ。

ケモ的には、PLS は「スペクトルの中でも、濃度や品質とちゃんと関係している“役に立つ”変動だけを優先して拾う手法」と言える。

5.2.2　アルゴリズム（概念的 PLS1 の流れ）

PLS1（応答が 1 つ）の典型的な流れは以下の通りである。

1. **共分散最大方向の特定**: まず大まかに $\mathbf{w} \propto \mathbf{X}^T\mathbf{y}$ から出発し、目的関数を満たす $\mathbf{w}_1$ を求める。潜在スコア $t_1 = \mathbf{X}\mathbf{w}_1$ を計算する。
2. **回帰**: $\mathbf{y}$ を t_1 で回帰して係数を得る。これは「成分 1 がどれくらい $\mathbf{y}$ を説明しているか」を意味する。
3. **デフレーション**: $\mathbf{X}$ と $\mathbf{y}$ から「t_1 によって説明された部分」を引き算して残差を作る。
4. **反復**: 残差データに対して同じ操作を繰り返し、必要な数だけ成分を抽出する。

デフレーションを行うことで、各成分が「重複の少ない、新しい情報」を持つようになる。

5.3　PCR と PLS のケモメトリクス的比較

表 3　PCR と PLS の比較

観点	PCR	PLS
軸の決め方	$\mathbf{X}$ の分散のみ	$\mathbf{X}$ と $\mathbf{y}$ の共分散
目的	スペクトルの再現性	応答の予測精度
HDLSS での安定性	○（次元削減済み）	○（次元削減＋応答利用）
欠点	$\mathbf{y}$ に重要だが分散が小さい方向を捨てる	アルゴリズムと成分数選択がやや複雑
ケモ的イメージ	「見た目の違いをよく説明」	「濃度・品質に効く違いを優先」

実務では、「まず PCA/PCR でざっくり構造を見る」、「本格的なキャリブレーションや判別には PLS (PLS-DA) を使う」という使い分けが多い。

6　PLS の発展的応用

前章で述べた PLS は、その強力な次元削減能力により、判別分析や変数選択といったより高度なタスクにも応用されている。本章では、代表的な応用例である PLS-DA とスパース PLS について解説する。

6.1　PLS 判別分析 (PLS-DA)

PLS-DA は、応答 $\mathbf{y}$ が連続値ではなく「クラスラベル」のときに使うバリエーションである。例えば 3 クラスの場合、各サンプルをダミーベクトル $\mathbf{Y} \in \mathbb{R}^{n\times 3}$ に変換し、これに対して PLS を行う。新しいサンプルを予測する際は、「もっとも予測値が大きいクラス」に分類する（分類ルール）。

ケモメトリクスでは、ガソリングレード（高オクタン vs 低オクタン）、食品の産地・品種、バッチの良品／不良品などの分類問題に対して、「**PCA+LDA よりも、多重共線性に強く、波長選択と相性が良い分類器**」として広く使われている。

6.2 スパース PLS (Sparse PLS) と変数選択

PLS は全波長を使って成分を作るため、「どの波長が重要か」が分かりにくい場合がある。これを補うのが VIP とスパース PLS である。

6.2.1 変数重要度 (VIP)

VIP (Variable Importance in Projection): 各波長が応答 $\mathbf{y}$ の説明にどれだけ貢献したかを 1 本のスコアにまとめた指標。通常、VIP > 1 の変数が重要とみなされる。

6.2.2 スパース PLS (SPLS)

スパース PLS: PLS の最適化に L_1 ペナルティを加え、重み $\mathbf{w}$ の多くの要素を**ちょうど 0 にする**。これにより、「本当に効いている波長だけに成分を支えさせる」ことができ、モデルの解釈性向上と波長数の削減（測定コスト削減）につながる。$\mathbf{w}$ の NIPALS 更新ステップにおいて、小さな重みは正確にゼロに設定される（ソフト閾値処理）。

パラメータチューニング:

- λ_1: スパース性を制御する（小：解釈性向上、大：特徴選択）。
- λ_2: リッジパラメータ（数値的安定性）。

これにより、メタボロミクスなどの高次元データにおける解釈性が大幅に向上する。

7 QSAR/QSPR と分子記述子

定量的構造活性相関 (QSAR) モデリングは、統計モデルを用いて化学構造と生物学的活性を関連付けるものである。

7.1 分子記述子

化学構造は数値ベクトル $\mathbf{x}_i \in \mathbb{R}^d$ に変換される。

- **0D/1D**: 原子数、分子量。
- **2D**: トポロジカルインデックス（結合性、パス長）。
- **3D**: 幾何学的および量子化学的記述子（構造最適化が必要）。

8 定量的構造活性相関 (QSAR/QSPR)

医薬品設計や材料科学において、分子構造からその活性や物性を予測する技術は QSAR（活性相関）または QSPR（物性相関）と呼ばれる。本質的には、分子構造を「数値」に変換し、それを入力として回帰モデルを作るプロセスである。

8.1 分子記述子：構造をどうやって「数字」にするか

QSAR/QSPR では、分子構造（構造式や 3D 構造）を計算機が扱える**数値ベクトル** $\mathbf{x}_i \in \mathbb{R}^d$ に変換する。これが分子記述子である。

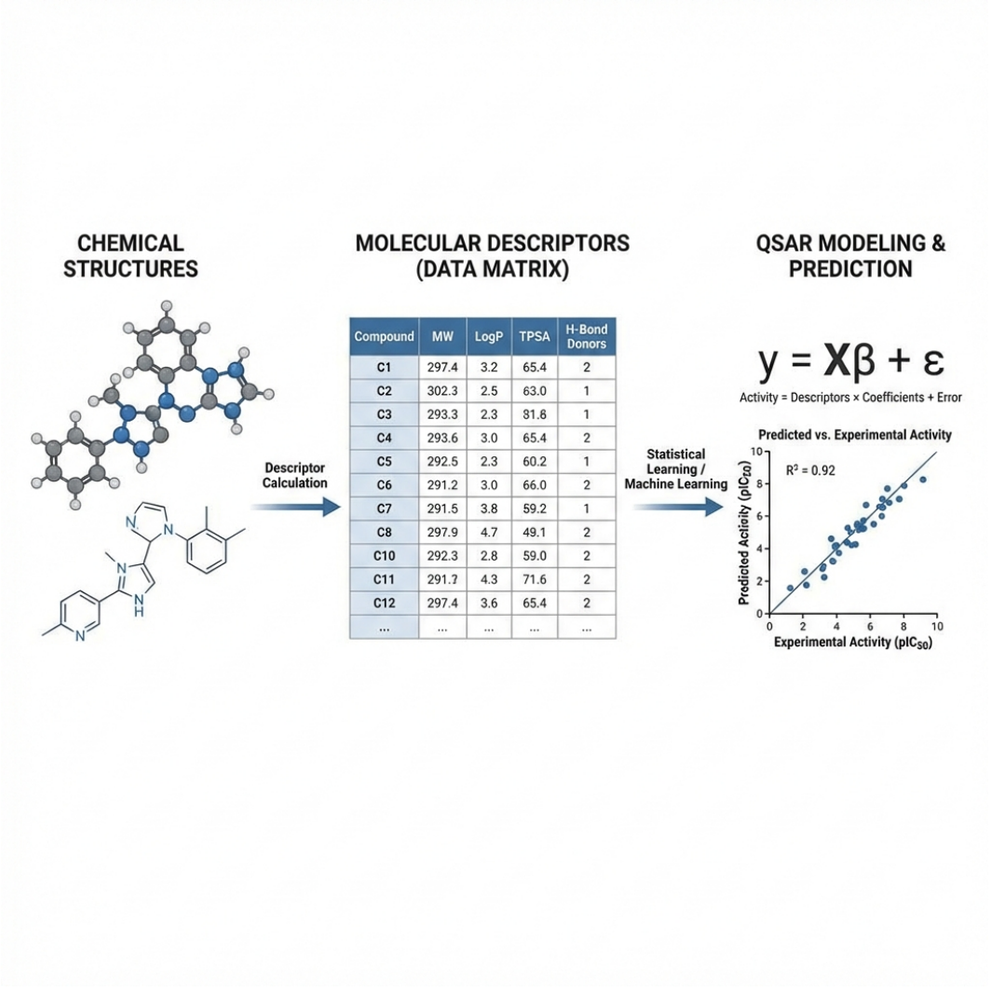

図 4　QSAR モデリングの概念的ワークフロー。化学構造は高次元の記述子ベクトルに変換され、生物学的活性にマッピングされる。**直観**: 核心となる考え方は「類似した構造 $\Longrightarrow$ 類似した活性」である。構造を数値（記述子）に定量化することで、未測定の化合物の生物学的効果を数学的に予測することができる。

8.1.1　0D / 1D 記述子（「どんな原子をどれくらい持つか」）

構造の細かいつながり方は無視し、**全体の数や性質**を表す記述子である。

- 原子数（C, H, N, O, F の個数）や分子量。
- 元素組成比（C/H 比、O/C 比など）。
- logP（オクタノール/水分配係数）：疎水性の指標。
- 極性表面積 (PSA)：極性官能基がどれくらい露出しているか。

直感的には、「その分子が重いか軽いか」「脂溶性か水溶性か」といった、**全体的なキャラクター**を表す。

8.1.2　2D 記述子（グラフとしての構造）

分子を「点（原子）と線（結合）」からなるグラフと見なし、そのつながり方のパターンを数値化したものである（トポロジカルインデックス）。

- **Wiener index**: 全ての原子間の距離（結合数）の総和。
- **Randic index**: 各結合の分岐度に基づく指標。
- 環の数（ベンゼン環など）や枝分かれ度。

これらは、「同じ原子数でも形が違う分子（異性体）を区別する」ために重要である。

8.1.3 3D 記述子（立体構造・量子化学的性質）

分子の 3D 構造を最適化したうえで、その構造や電子状態から計算する記述子である。

- **幾何学的記述子**: 分子の体積、表面積、慣性モーメント（形状の広がり）。
- **量子化学記述子**: HOMO/LUMO エネルギー（反応性）、双極子モーメント（電荷の偏り）、Fukui 関数（局所反応性）。

3D 記述子は、立体配座や電子状態が活性に強く効くケース（受容体への結合、光物性など）で特に重要となる。

8.2 PLS-QSAR モデル

多数の分子記述子 $\mathbf{x}_i$ (d 次元) と、その活性・物性 y_i（IC50, 溶解度など）を結びつけるために PLS が用いられる。

PLS では、まず重み $\mathbf{w}_a$ と潜在スコア $t_{ia} = \mathbf{x}_i^T \mathbf{w}_a$ を求める。活性 y_i は以下のようにモデル化される：

$$y_i = b_0 + \sum_{a=1}^{A} c_a t_{ia} + \varepsilon_i \tag{4}$$

- t_{ia}: 分子 i の「潜在変数 a 上の位置」。多数の記述子をまとめた「構造パターン」。
- c_a: その潜在変数が活性にどれくらい効いているか。

ケモ的な意味: 第 1 成分 t_1 は「活性と最も共分散が大きい構造パターン」（例：疎水性＋芳香族性の組み合わせ）、第 2 成分 t_2 は「残りの部分で活性と関係するパターン」を表す。これにより、記述子数が分子数より多い HDLSS 状況や、記述子間の多重共線性があっても安定してモデルを構築できる。

8.3 モデル検証統計量：R^2 と Q^2 の意味

QSAR/QSPR では、単なる当てはまりだけでなく「外れたデータに対する予測能力」が必須である。

8.3.1 決定係数 R^2（当てはまりの良さ）

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{5}$$

R^2 は「訓練データの説明力」であり、パラメータを増やせば 1 に近づくため、単独では過学習の可能性を排除できない。

8.3.2 交差検証係数 Q^2（予測能力）

交差検証 (LOO や k-fold) において、各分子 i を「予測専用」に回したときの予測値 $\hat{y}_{i,\mathrm{CV}}$ を用いて計算する。

$$Q^2 = 1 - \frac{\sum (y_i - \hat{y}_{i,\mathrm{CV}})^2}{\sum (y_i - \bar{y})^2} \tag{6}$$

Q^2 は「未知データに対する予測力」を表す指標である。

8.3.3 OECD の受容基準と過学習チェック

典型的な受容基準は以下の通りである。

- $R^2 > 0.8$: 訓練データへの十分な当てはまり。

- $Q^2 > 0.6$: 交差検証でも十分な予測性能。
- $R^2 - Q^2 < 0.1$: 過学習が小さいこと。

もし $R^2 \gg Q^2$（例：$R^2 = 0.95, Q^2 = 0.3$）の場合、モデルは訓練データのノイズを「暗記」しているだけであり、新しい化合物には通用しない（過学習）。したがって、(1) R^2 で説明力を確認し、(2) Q^2 で予測力を検証し、(3) 両者の差で過学習を判断する、という 3 段階チェックが必須である。

9 HDLSS 環境における統計理論

9.1 この章で扱うこと（導入）

これまでの章では、PCA や PLS を「使い方」と「化学的な解釈」の観点から説明してきた。しかし、スペクトルやマイクロアレイのような **HDLSS (High-Dimension, Low-Sample-Size, $d \gg n$)** 環境では、これらの手法は古典的な前提 ($n \to \infty$ で d は固定) から大きく外れた状況で使われている。

その結果として、次のような現象が起こることが知られている。

- 距離や共分散に基づく「直感」が崩れる（測度の集中、次元の呪い）。
- 標本共分散行列の固有値・固有ベクトルが、母集団のものから系統的にずれる（不一致性）。
- PCA のスクリー図や、PLS の成分数の選択が、「見た目の印象」に強く依存してしまう。

この章では、青嶋・矢田らによって発展してきた HDLSS 統計理論、とくに

- **強スパイク固有値 (Strong Spiked Eigenvalue, SSE) モデル**
- **ノイズ削減 (Noise Reduction, NR) 法**

を紹介し、それらが **ケモメトリクスにおける PCA/PLS の解釈とチューニングにどのような理論的裏付けを与えるか** を概観する。数学的な詳細や証明は **付録 付録 D** に譲り、本章では「どの式が、どのような幾何学的／化学的意味を持つか」に焦点を当てる。

9.2 強スパイク固有値 (SSE) モデル

9.2.1 シグナル＋ノイズとしての共分散構造

母共分散行列 Σ の固有値を $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ と並べる。SSE モデルでは、次のような構造を仮定する。

- 上位 k 個の固有値 $\lambda_1, \ldots, \lambda_k$ は、次元 d の増大とともに $\lambda_j \sim d^{\alpha}$ ($\alpha \geq 0.5$) のオーダーで大きくなる「強いスパイク」である。
- それ以降の固有値は、比較的小さく、全体として「ノイズ部分空間」の分散を構成する。

ケモメトリクス的解釈:

- $\lambda_1, \ldots, \lambda_k$: **主要な化学成分の濃度変動**や、支配的な物理変動（例えば水分・脂肪・温度）の寄与。
- 残りの多数の固有値: 機器ノイズ、微弱な成分、前処理で取り切れなかったベースライン揺らぎなどから成る「バルクノイズ」。

SSE モデルは、「少数の強いシグナルと多数の弱いノイズが共存する」という、現実の高次元ケモメトリクスデータをよく表す理想化モデルである。

9.2.2 測度の集中と距離の直感の崩壊

SSE モデルのもとで、標本平均 $\bar{\mathbf{x}}$ と真の平均 $\boldsymbol{\mu}$ の距離は以下の式を満たす。

$$\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|^2 = \frac{\mathrm{tr}(\Sigma)}{n}\{1 + o_P(1)\} \tag{7}$$

この式は、次元 d が大きくなるにつれ、距離が「ほぼ決まった半径」に集中してしまうことを意味する。

幾何学的直感: 高次元では、すべての点が平均から**ほぼ同じ距離**だけ離れる。その結果、点と点の距離も「ほとんど同じ」に見えやすくなる（測度の集中）。

ケモメトリクス的な含意: ユークリッド距離に基づく k-NN などの分類器は、HDLSS では「どのサンプルも同じような距離」に見えてしまい、ノイズ構造に強く左右される。同様に、PCA のスコア空間でも、「距離」や「角度」に対する通常の直感が当てはまらなくなる場合がある。これが、HDLSS で PCA/PLS を使う際に「距離ベースの解釈」に慎重さが必要な理由である。

9.3 ノイズ削減 (NR) 法

9.3.1 標本 PCA 固有値のバイアス

標本共分散行列 $\mathbf{S}$ の固有値を $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$ とする。SSE モデルのもとでは、通常の PCA から得られる $\hat{\lambda}_j$ は、母固有値 λ_j に対して系統的なバイアスを持つことが知られている。「シグナル＋ノイズ」の構造の下では、シグナル固有値はノイズ部分空間の分散によって**押し上げられる**。単純なスクリー図では、この「押し上げ」と「揺らぎ」が混ざるため、どこまでが有意な成分かの判断が主観的になりがちである。

9.3.2 NR 推定量の定義と化学的解釈

矢田・青嶋によるノイズ削減 (NR) 法は、標本固有値を「高次元ノイズトレース」によるバイアスから補正する枠組みである。第 j 固有値の NR 推定量 $\tilde{\lambda}_j$ は以下で与えられる（詳細は付録 付録 D 参照）。

$$\tilde{\lambda}_j = \hat{\lambda}_j - \frac{\mathrm{tr}(\mathbf{S}_D) - \sum_{s=1}^{j} \hat{\lambda}_s}{n - 1 - j} \tag{8}$$

- 第一項 $\hat{\lambda}_j$: 観測された「シグナル＋ノイズ」の固有値。
- 第二項（減算項）: それ以降の多数の成分が持つ「バルクノイズの平均レベル」。

ケモメトリクス的解釈: $\tilde{\lambda}_j$ は、「第 j 成分に対応する**純粋な化学シグナルの強さ**」を推定しているとみなせる。これにより、PCA の固有値に対して「どの成分までが実質的なシグナルで、どこから先がノイズか」をより客観的に判別しやすくなる。

9.3.3 SNR に基づく成分選択

さらに、推定されたノイズ分散 $\hat{\kappa}$ を用いて、成分ごとの S/N 比を定義する。

$$\mathrm{SNR}_j = \frac{\tilde{\lambda}_j}{\hat{\kappa}} \tag{9}$$

一般的な経験則として、$\mathrm{SNR}_j > 3$ ならば有意なシグナル成分、$\mathrm{SNR}_j \leq 3$ ならば主にノイズとみなす。ケモメトリクス的には、これは PCA や PLS の成分数選択において、従来のヒューリスティックに加えて「NR 補正後の SNR に基づく**理論的根拠を持った選択**」を提供するものである。

10 実装とソフトウェア環境

本章では、R を用いたケモメトリクスパイプラインの実装について実践的なガイドを提供する。化学ベンチマークデータセットを厳密に使用する。

10.1　再現可能なコード資産

本稿で提示された解析を再現するための完全な実行可能スクリプトを提供する。これらのスクリプトはプロジェクトの `code/` ディレクトリにある。

- `code/01_data_export.R`: `pls` パッケージから `gasoline` データセットをエクスポートする。
- `code/02_pipeline.R`: 以下の図を生成する完全な R ワークフロー。
- `code/03_pipeline.py`: 同等の Python ワークフロー。

10.2　実践例：ガソリン分析

`gasoline` データセット（$n = 60$, $d = 401$, NIR スペクトル）を解析する。以下を実演する：

1. **回帰**: オクタン価（定量的）の予測。
2. **分類 (PLS-DA)**: 高オクタン価グレードと低オクタン価グレードの識別（定性的）。

10.2.1　1. データの可視化

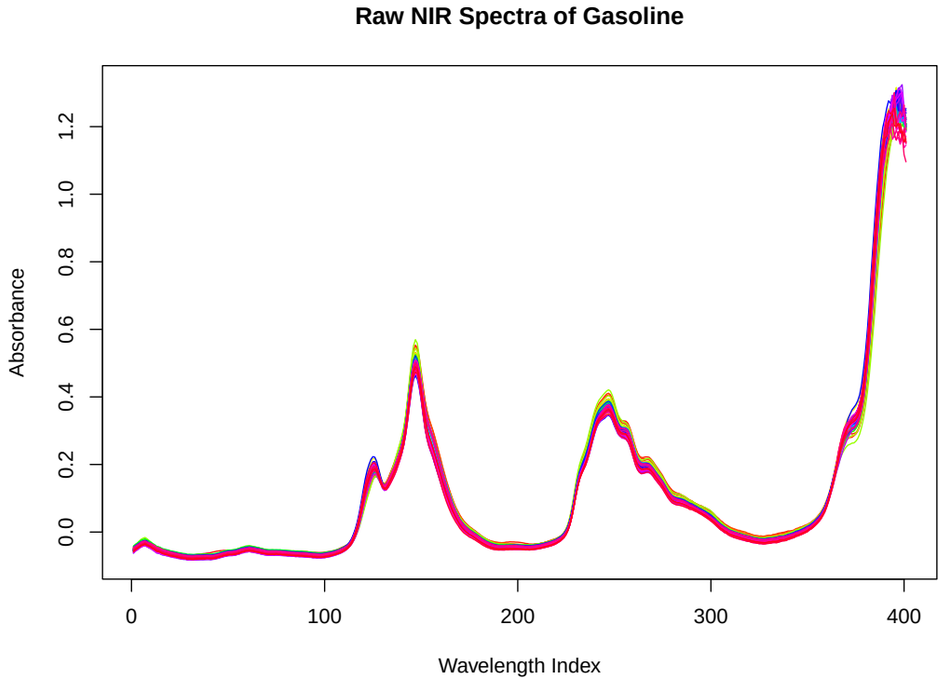

図 5　60 個のガソリンサンプルの生 NIR スペクトル。**直観**: 各線は化学的指紋である。重なり合うピークは、異なる燃料成分（ヘプタン、イソオクタンなど）の振動する C-H 結合に対応する。我々の目標は、この複雑な重なりを解読してオクタン価を予測することである。

10.2.2　2. 回帰 (RMSEP 分析)

オクタン価を予測するために PLS モデルをキャリブレーションする。

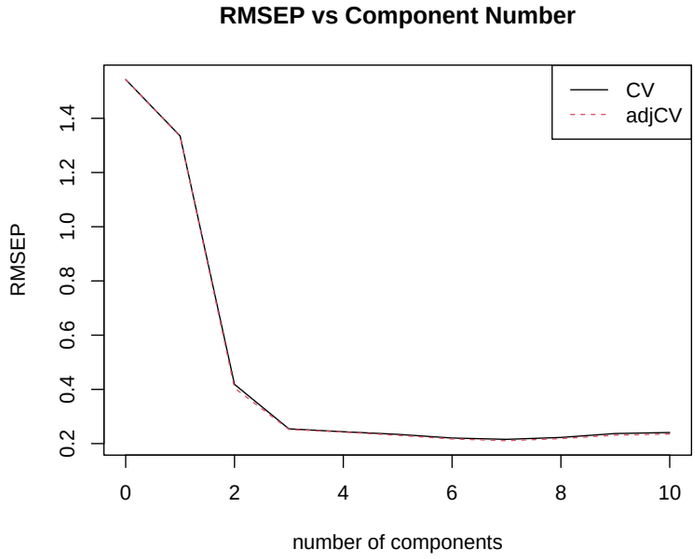

図 6　RMSEP 対成分数。有益な成分を追加すると誤差は本質的に低下し、ノイズを追加すると（過学習）横ばいになるか上昇する。「膝（knee）」が生じる成分数（ここでは通常 2-3）を選択し、精度と単純さのバランスをとる。

10.2.3　3. 分類 (PLS-DA)

連続的なオクタン価を「高オクタン」（> 87）と「低オクタン」（≤ 87）の 2 値クラスに変換する。その後、PLS-DA を適用して潜在空間での分離を可視化する。

11　ケーススタディ

11.1　ケース 1: ビールの NIR 分析 (iPLS)

データセット: $n = 60$ ビールサンプル, $d = 256$ 波長 (NIR)。応答 y: 原麦汁エキス濃度 (°Plato)。
標準的な PLS の結果:

- 潜在変数 (LV): 5
- RMSECV: 0.24°Plato
- R^2_{CV}: 0.994

インターバル PLS (iPLS) 最適化: スペクトルを 40 の区間に分割した。最適な領域は 1150-1500 nm ($d_{opt} = 50$) であることがわかった。

- **RMSECV**: 0.17°Plato (30% 改善)
- **変数**: 256 から 50 に削減 (80% 削減)。
- **解釈**: 選択された領域は水/アルコールの O-H 倍音帯に対応しており、化学的に妥当である。

11.2　ケース 2: ガソリングレード分類 (PLS-DA)

データセット: $n = 60$ サンプル, $d = 401$ NIR 波長。**目的**: サンプルを「高オクタン」（> 87）グレードと「低オクタン」グレードに分類する。

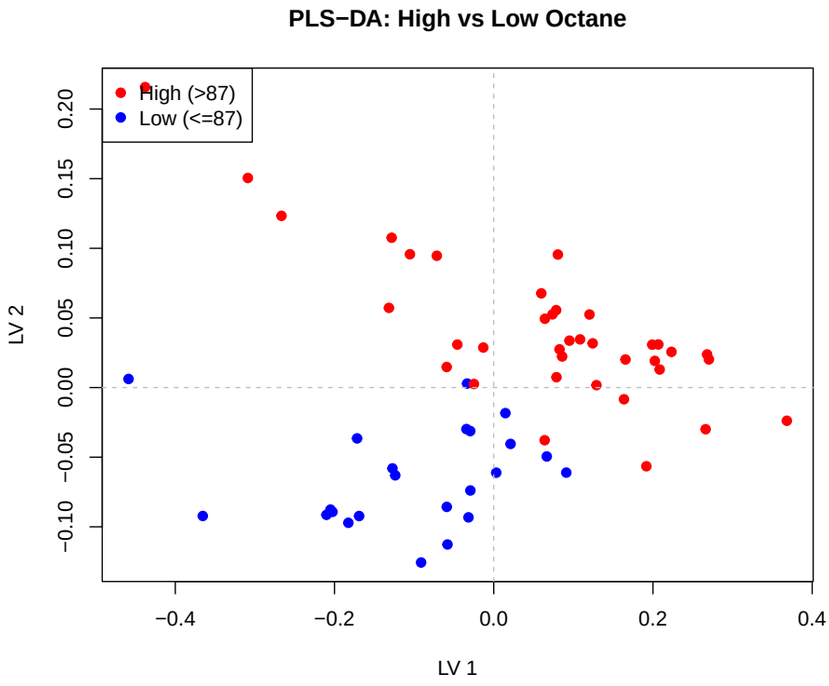

図 7　PLS-DA スコアプロット。赤：高オクタン、青：低オクタン。**直観**: この 2D マップは 401 次元の問題を単純化している。赤と青のクラスターが水平軸（成分 1）に沿って離れているという事実は、モデルが高グレード燃料と低グレード燃料を区別する化学的「レシピ」を見つけ出したことを意味する。

手法:

- **特徴抽出**: 2 つの潜在変数を持つ PLS-DA。
- **バリデーション**: 10 分割交差検証。

結果:

- **可視化**: スコアプロットは LV1 に沿った線形分離可能性を示す。
- **精度**:わずか 2 成分で 100% の分類精度を達成。
- **解釈**: ローディングは、芳香族（ブースター）対脂肪族に典型的な C-H 倍音振動に対応している。

これは、ケモメトリクス的パターン認識がいかにして製油所の品質管理 (QC) を自動化できるかを実証している。

12　今後の展望と課題：研究ロードマップ

本章では、ケモメトリクスの最前線における 3 つの主要な研究領域（テンソル分解、深層学習、因果推論）について、実際の研究で使えるレベルの数理モデルと具体的な研究テーマを整理する。これらは、単なる手法の適用にとどまらず、化学的知見とデータサイエンスの深い融合が求められる領域である。

12.1　テンソル分解（マルチウェイ解析）

12.1.1　データ構造と記法

三元配置データ（例：励起蛍光マトリックス EEM、時間分解 NIR、GC×GC-MS など）は、以下の 3 次テンソルとして表される。

$$\mathcal{X} \in \mathbb{R}^{I \times J \times K} \tag{10}$$

ここで、$i = 1, \ldots, I$ はサンプル、$j = 1, \ldots, J$ は変数 1（波長や保持時間 1）、$k = 1, \ldots, K$ は変数 2（励起波長や保持時間 2）を表す。行列表現とは異なり、「サンプル × 波長 × 時間」といった**構造（連続性・位置関係）を保ったまま**解析できる点が最大の特徴である。

以下に、代表的なマルチウェイ測定とそのデータ構造を整理する。

■1. 励起蛍光マトリックス (EEM)

- **何を測っているか**: 分子に様々な波長の光（励起波長）を当て、そこから出てくる蛍光の波長と強度を測定したもの。「刺激する色」と「光る色」の組み合わせごとの蛍光強度マップ（指紋）である。
- **データ構造**: I（試料）× J（励起波長）× K（蛍光波長）。
- **解析タスク**: 複数の蛍光種が重なっているため、PARAFAC を用いて「成分ごとの濃度 × 励起スペクトル × 蛍光スペクトル」に分離し、定性・定量を行う。

■2. 時間分解 NIR (Time-resolved NIR)

- **何を測っているか**: 反応や乾燥プロセスにおいて、NIR スペクトルが時間とともにどう変化するかを追跡したもの。
- **データ構造**: I（試料/実験条件）× J（時間）× K（波長）。
- **解析タスク**: 反応中間体のスペクトルと生成プロファイルの抽出、あるいは N-PLS を用いた「時間帯 × 波長帯」の重要度評価。

■3. GC×GC-MS

- **何を測っているか**: 性質の異なる 2 本のカラムを直列につなぎ、成分を「保持時間 1」と「保持時間 2」の 2 軸で分離し、さらに MS で質量スペクトルを取得したもの。
- **データ構造**: I（試料）× J（RT1）× K（RT2）× L（m/z）。実務的には m/z を縮約して 3 次テンソルとして扱うことが多い。
- **解析タスク**: 複雑な共溶出成分を、2 次元クロマトグラム上のピーク形状と質量スペクトルに基づいて分離・同定する。

12.1.2　PARAFAC（並列因子分析）

典型的な 3 次 PARAFAC モデルは以下のように書かれる。

$$x_{ijk} \approx \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \tag{11}$$

- F: 成分数（潜在因子の数）
- $\mathbf{A} \in \mathbb{R}^{I \times F}, \mathbf{B} \in \mathbb{R}^{J \times F}, \mathbf{C} \in \mathbb{R}^{K \times F}$: 各モードのローディング行列
- e_{ijk}: 残差

行列表現を用いると、Khatri-Rao 積 $\odot$ を用いて $\mathbf{X}_{(1)} \approx \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T$ と記述できる。

数学的性質と研究テーマ:

- **一意性:** PCA と異なり、PARAFAC は「回転の不定性」を持たず、軽い条件の下でパラメータが一意に定まる（Kruskal 条件）。これは、混合スペクトルから純粋な化学成分のプロファイルを分離できることを意味する。
- **ケモメトリクスでの応用:**
  - EEM データからの蛍光種スペクトル・励起プロファイル・濃度プロファイルの同時推定。
  - GC×GC-MS のような複雑なデータからの共溶出ピークの分解。

研究課題の例:

- 保持時間シフトやピーク形状の変化を許容する PARAFAC2 モデルの拡張。
- HDLSS（サンプル少・モード多）環境下での PARAFAC の推定誤差解析と、スパース正則化（L1 ノルム等）の導入。

12.1.3 N-PLS（マルチウェイ PLS）

三元テンソル $\mathcal{X}$ と一次元応答 $\mathbf{y} \in \mathbb{R}^I$ の回帰問題に対し、N-PLS は各コンポーネント f ごとにスコア $\mathbf{t}_f$ とローディング $\mathbf{w}_f^{(1)}, \mathbf{w}_f^{(2)}$ を求める。

$$t_{if} = \sum_{j=1}^{J} \sum_{k=1}^{K} x_{ijk} w_{jf}^{(1)} w_{kf}^{(2)} \tag{12}$$

目的関数の一例は以下の通りである。

$$\max_{\mathbf{w}_f^{(1)}, \mathbf{w}_f^{(2)}} \mathrm{cov}^2(\mathbf{t}_f, \mathbf{y}) \quad \text{s.t.} \quad ||\mathbf{w}_f^{(1)}|| = ||\mathbf{w}_f^{(2)}|| = 1 \tag{13}$$

研究課題の例:

- PARAFAC ベースの N-PLS に対する HDLSS 理論（固有値バイアス補正など）。
- スパース N-PLS：各モードのローディングに L1 正則化を課し、時間帯 × 波長帯のブロック選択を行う高次元テンソル回帰。

12.2 ケモメトリクスにおける深層学習

12.2.1 1D-CNN によるスペクトル表現学習

スペクトル $\mathbf{x} \in \mathbb{R}^d$ に対し、1 次元畳み込み層は以下で定義される。

$$h_k^{(1)}(i) = \sigma\left(\sum_{m=1}^{M} w_{k,m}^{(1)} x_{i+m-1} + b_k^{(1)} \right) \tag{14}$$

直感的には、「ローカルな波長帯の形状（ピークの形・幅・肩）」を自動的に特徴量として抽出する操作である。層を重ねることで、より高次のパターン（ピークの組み合わせ、基底バンド構造など）を学習する。

最終的に得られる特徴ベクトル $\mathbf{z}$ に対し、回帰 $\hat{y} = \mathbf{w}^T\mathbf{z} + b$ や分類 $\hat{\mathbf{p}} = \mathrm{softmax}(\mathbf{W}\mathbf{z} + \mathbf{b})$ を行うことで、従来の PLS/QSAR と同様の枠組みで出力を得る。

12.2.2 HDLSS での課題：パラメータ数と汎化

1D-CNN のパラメータ数は膨大になりがちであり、HDLSS（例：$n \sim 50$）では「訓練誤差はほぼ 0、テスト誤差は巨大」という過学習が顕著になる。PLS が「低次元の潜在空間」に射影してから回帰していたのに対

し、CNN は射影空間も含めて学習するため、統計的な制御が難しい。

12.2.3 転移学習と少数ショット学習

この問題に対する有望なアプローチが転移学習である。

1. 大規模スペクトルデータセット（例：大量の原材料スペクトル）で CNN を事前学習し、パラメータ $\theta^\star$ を得る。
2. 小規模なターゲットデータセットに対して、低層（基底フィルタ）は固定し、高層のみ微調整 (fine-tuning) する。

これは数学的には、パラメータ空間の探索を「事前学習で得た良い解の近傍」に制限することと解釈でき、有効パラメータ数を実質的に減らす効果がある。

研究課題の例:

- PLS の潜在空間と CNN 中間層の表現を整合させる「PLS-正則化 CNN」。
- SSE モデル・ノイズ削減 (NR) 法の観点から、CNN が生成する特徴空間の「スパイク固有値構造」を解析し、HDLSS における汎化誤差境界を導出すること。

12.3 因果推論

12.3.1 PLS はあくまで「相関」を捉える

PLS 回帰は相関が高い方向を探す手法であり、「ある波長帯の吸光度が高いと品質が良い」という関係は見つけられるが、「その波長帯が原因で品質が良くなった」とは断言できない。温度や原材料ロットなどの未観測の交絡因子が、見かけ上の相関を生んでいる可能性がある。

12.3.2 構造方程式モデリング (SEM)

SEM では、観測変数ベクトル $\mathbf{z}$（スペクトル特徴、プロセス条件、応答など）が以下の線形構造に従うと仮定する。

$$\mathbf{z} = \mathbf{B}\mathbf{z} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\delta} \tag{15}$$

- $\mathbf{B}$: 観測変数間の因果パス（隣接行列）。
- $\boldsymbol{\xi}$: 潜在変数（「プロセスの健全性」「配合設計」など）。
- $\boldsymbol{\Gamma}$: 潜在変数から観測変数への影響。

得られたモデルの適合度を評価することで、「プロセス条件 $\rightarrow$ 中間品質（NIR スコア） $\rightarrow$ 最終品質」といった因果的メカニズムを検証できる。

12.3.3 ベイジアンネットワーク (BN) と介入

BN は、有向非巡回グラフ (DAG) 上の確率モデルであり、学習されたグラフから「介入 ($do(\cdot)$ 演算)」の効果を計算できる。これにより、「もし温度を意図的に固定したら、スペクトルと品質はどう変化するか」といった因果効果を予測できる。

研究課題の例:

- スペクトル特徴（PLS スコア）とプロセス条件をノードとする BN を構築し、因果パスを可視化する。
- SEM と BN の統合フレームワークを化学プロセスデータに適用し、相関分析を超えた「メカニズムの解明」を目指す。

付録 A　ベンチマークデータセット

本付録では、ケモメトリクスの学習や研究で標準的に使用されるベンチマークデータセットについて詳述する。これらは単なる数値の羅列ではなく、それぞれが固有の化学的・統計的課題（装置差、HDLSS、非線形性など）を含んだ優れた「教材」である。

A.1　代表的な分光データセット

多くは R パッケージ（`pls` など）に同梱されており、PCA、PLS、前処理、波長選択のデモに即座に使用可能である。

A.1.1　トウモロコシ (Corn) データセット

- **データ構造**: 3 台の異なる NIR 分光計（m5, mp5, mp6）で、同じ種類のトウモロコシ試料（80 サンプル）を測定したデータ。波長範囲は 1100–2498 nm（2 nm 間隔）。
  - 入力変数: 各波長における吸光度。装置ごとに微妙なオフセットや感度差（装置差）が含まれるのが特徴。
- **応答変数（解析タスク）**: 水分、油分、タンパク質、デンプンの 4 成分の含有量。これらを同時に予測するマルチ応答 **PLS 回帰 (PLS2)** の典型的な教材である。
- **教育的・実務的価値**: 最大の特徴は「キャリブレーション移送」のベンチマークである点だ。ある装置（m5）で構築したモデルを、追加実験なしで別の装置（mp6）に適用できるか？という、Piecewise Direct Standardization (PDS) やドメイン適応（Transfer Learning）の実験に最適である。

A.1.2　ガソリン (Gasoline) データセット

- **データ構造**: 60 サンプル、401 波長（900–1700 nm）の NIR スペクトル。サンプル数 ($n = 60$) よりも変数 ($d = 401$) が多い、典型的な **HDLSS** ($n \ll d$) データである。
  - 入力変数: ガソリン中の炭化水素（ヘプタン、イソオクタン等）の C-H 結合振動に由来する吸光度。
- **応答変数（解析タスク）**: オクタン価（単一応答）。ガソリンの「燃えにくさ（ノッキング耐性）」を表す指標であり、高いほど高品質とされる。基本タスクは **PLS1 回帰**である。
- **教育的・実務的価値**: 変数が多いため、通常の重回帰分析 (OLS) は破綻する。PLS がいかにして多重共線性を克服するかを示すのに適している。また、iPLS や遺伝的アルゴリズムによる**波長選択**を行い、「全波長を使うより、重要な波長だけ選んだ方が精度が上がる」現象（スパース性）を体験させるのにもよい。

A.1.3　Tecator 食肉 (Meat) データセット

- **データ構造**: 215 個の肉サンプルの NIR スペクトル（850–1050 nm）。
- **応答変数（解析タスク）**: 脂肪、水分、タンパク質の含有量。特に**脂肪含有量**はスペクトルとの関係が**非線形**であることが知られている。
- **教育的・実務的価値**: 「線形モデル（PLS）の限界」を示す教材として重要である。PLS では高脂肪領域で予測誤差（バイアス）が残るが、ニューラルネットワーク (ANN) やカーネル PLS (KPLS) などの非線形モデルを導入すると劇的に改善する。この対比を通じて、モデル選択の重要性を学べる。

A.2 その他の有名なデータセット

A.2.1 Wine データセット

- **データ概要**: イタリアの同一地域で生産された 3 品種のワインについて、13 種の化学成分（アルコール度数、リンゴ酸、灰分、フラボノイド、色強度など）を測定したデータ ($n = 178, d = 13$)。
- **用途**: **PCA による可視化**の入門に最適。スペクトルではなく「意味のある化学値」が変数であるため、ローディングプロット（どの成分が品種の違いに寄与しているか）の解釈が容易である。LDA や PLS-DA による分類タスクにも使われる。

A.2.2 Olive Oil データセット

- **データ概要**: 異なる産地（イタリアの複数の地域）のオリーブオイルについて、8 種類の脂肪酸（オレイン酸、リノール酸など）の組成比率を測定したデータ ($n = 572, d = 8$)。
- **用途**: 産地判別（地理的プロファイリング）。地理的な階層構造（南部、サルデーニャ島、北部など）が PCA プロット上で綺麗に分かれるため、クラスター分析や教師あり学習のデモに適している。

A.2.3 Tablets (錠剤) データセット

- **データ概要**: 製薬プロセスにおける錠剤の NIR スペクトル ($n = 310, d = 404$)。
- **用途**: **PAT (Process Analytical Technology)** の実例。スペクトルから有効成分 (API) の含有量を非破壊で瞬時に推定する PLS モデルの構築に使われる。「製造ラインでのリアルタイム全数検査」を模擬できる。

A.2.4 QM9 (量子化学) データセット

- **データ概要**: 約 13 万個の小分子有機化合物について、量子化学計算（DFT）で求めた物性値（HOMO/LUMO エネルギー、双極子モーメント、熱容量など）をまとめた大規模データセット。
- **用途**: **深層学習 (Deep Learning)** のベンチマーク。入力はスペクトルではなく「分子グラフ（原子のつながり）」や「3D 座標」である。従来の QSPR (構造物性相関) と、最新のグラフニューラルネットワーク (GNN) の性能比較によく用いられる。

A.3 パブリックデータの比較と活用指針

表 4 に各データセットの特徴を整理した。「どのデータセットで、何を学ぶか」の指針として活用されたい。

A.4 データセットの入手方法

- **R / Python ライブラリ**: `pls` パッケージ (R) には `gasoline`, `yarn`, `oliveoil` 等が含まれる。`caret` や `scikit-learn` 経由でアクセス可能なものも多い。
- **リポジトリ**: Eigenvector Research (Corn, Tablets 等) や UCI Machine Learning Repository (Wine 等) が主要なソースである。

表 4　代表的なケモメトリクス・ベンチマークデータセットの比較と活用テーマ

データセット	データタイプ	主な解析タスク	想定する授業・研究テーマの例
Corn	NIR スペクトル	マルチ応答回帰 (PLS2)	装置差補正、ドメイン適応、キャリブレーション移送
Gasoline	NIR スペクトル	単一応答回帰 (PLS1)	HDLSS 入門、波長選択 (iPLS, GA)、過学習の理解
Tecator	NIR スペクトル	非線形回帰	線形 (PLS) vs 非線形 (ANN/SVM) の比較、残差分析
Wine	化学組成	分類、探索的解析	PCA/LDA による可視化、因子負荷量の解釈
Olive Oil	脂肪酸組成	分類 (産地判別)	地理的プロファイリング、階層的クラスター構造
Tablets	NIR スペクトル	回帰 (API 定量)	PAT（プロセス分析）、製造バッチ管理
QM9	分子構造 (グラフ)	構造物性相関 (QSPR)	深層学習 (GNN)、ビッグデータ解析、分子設計

付録 B　標準的な解析ワークフロー

堅牢なケモメトリクス解析には体系的なアプローチが不可欠である。標準的なワークフローは、5 つの主要なフェーズで構成される。前処理でノイズを消し、この流れで解析を行うことで、複雑な化学データから真実を見つけ出すことができる。

B.1　フェーズ 1: データの構造を知る（探索的データ解析）

まずはデータの中にどのようなパターンやグループがあるかを、先入観なしに調べる。

- **主成分分析 (PCA)**: 数百ある波長のデータを、情報の重要度が高い順に数本の「主成分」に凝縮する。
- **スコアプロットの確認**: 凝縮したデータをグラフにプロットし、サンプルが似たもの同士で集まっているか、明らかに異常なデータ（外れ値）がないかを目視でチェックする。

B.2　フェーズ 2: サンプルの分割（トレーニングとテスト）

作成したモデルの「本当の実力」を測るために、データをあらかじめ 2 つに分ける。

- **トレーニングセット (約 2/3)**: モデルを作るために使うデータ。
- **テストセット (約 1/3)**: モデルが完成するまで隠しておき、最後に「抜き打ちテスト」をするためのデータ。

B.3　フェーズ 3: モデルの構築（キャリブレーション）

トレーニングデータを使って、予測のための計算式を作る。

- **PLS（部分的最小二乗法）**: スペクトルの変動と、知りたい値（濃度など）の相関が最大になるように、最適な重み付けを計算する。
- **潜在変数の決定**: 情報の凝縮をどこまで行うか（主成分の数など）を、交差検証 (Cross-validation) などを用いて慎重に決定する。凝縮しすぎると情報不足になり、しなさすぎるとノイズまで学習（過学習）

してしまう。

B.4　フェーズ 4: モデルの検証（バリデーション）

完成したモデルが、新しいサンプルに対しても正しく動くかを確認する。

- **外部バリデーション**: 隠しておいた「テストセット」をモデルに入れ、予測値と実際の値のズレ（RMSEP: 予測の二乗平均平方根誤差）を計算する。
- **決定係数 (R^2)**: 予測値と実測値がどれくらい一致しているかを 0〜1 の数値で評価する（1 に近いほど優秀）。

B.5　フェーズ 5: 解釈と応用

最後に、なぜその結果になったのかを化学的に納得できるか確認する。

- **ローディングプロット/VIP スコア**: 「AI はどの波長を重要だと判断したか」を調べる。その波長が、対象物質の特定の化学結合（C-H 結合など）と一致していれば、そのモデルは信頼できると判断される。

表 5　解析の流れまとめ

ステップ	主な手法	目的
1. 探索	PCA	データの全体像と異常値の把握
2. 分割	サンプリング	公平な評価のための準備
3. 構築	PLS / PCR	濃度などを当てる計算式を作る
4. 検証	RMSEP / Q^2	モデルの実力を数値で測る
5. 解釈	係数分析	化学的な根拠を裏付ける

付録 C　ケモメトリクスのための数学的準備

本付録では、化学的なバックグラウンドを持つ読者（高校生〜学部初年級）を対象に、本稿で使用される数学的概念と記法について、「なぜそれが必要なのか」「化学的にどういう意味があるのか」を含めて解説する。

C.1　行列とベクトルの記法（もう一歩ていねいに）

C.1.1　データを「表」として見る

ケモメトリクスでは、データを以下のような「表（行列）」として扱うのが標準的である。

- **スカラー** (x): ただの 1 つの数（例：あるサンプルの脂肪含有量 12.3%）。
- **ベクトル** ($\mathbf{x}$): 数が縦（または横）に並んだもの。

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \tag{16}$$

  ここで x_j は「波長 j での吸光度」を表す。
- **行列** ($\mathbf{X}$): ベクトルを並べた「表」。

- 行 **(Row)**: サンプル（試料）。「1 行」は「1 つのサンプルの全スペクトル」に対応する。
- 列 **(Column)**: 変数（波長）。「1 列」は「特定の波長における全サンプルの吸光度」に対応する。

化学データではほぼ常に「行＝サンプル、列＝変数」という約束にしておくと、PCA などの式が直感的に理解しやすくなる。

C.1.2 添字の意味を明記する

- x_{ij}: i 行 j 列の要素。「i 番目のサンプルの、j 番目の波長での吸光度」。
- $\mathbf{x}_i$: i 番目のサンプルのスペクトル（1 行分を取り出したベクトル）。
- $\mathbf{x}^{(j)}$: j 番目の波長における全サンプルの値（1 列分を取り出したベクトル）。

このイメージを持つことで、PCA の式 $\mathbf{X}_c\mathbf{p}_j = \mathbf{t}_j$ が、「データ表 $\mathbf{X}_c$ に 重み $\mathbf{p}_j$ を掛けて、新しい指標 $\mathbf{t}_j$ を作る」操作であることが見えてくる。

C.2 行列形式での基本統計（直感補強）

C.2.1 平均ベクトルの意味

平均スペクトル $\bar{\mathbf{x}}$ は、数式では $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}^T\mathbf{1}_n$ と書かれるが、これは単に**「各波長ごとの平均値を計算して並べたもの」**に過ぎない。

- **スカラー表現**: $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ （波長 j の平均）

化学的な意味: 「平均スペクトル」は、そのデータセット全体に共通する「典型的な試料の姿」を表す。ベースライン補正や MSC などの前処理では、この平均スペクトルが「基準（Standard）」として頻繁に使用される。

C.2.2 中心化の幾何学的イメージ

中心化 $\mathbf{X}_c = \mathbf{X} - \mathbf{1}_n\bar{\mathbf{x}}^T$ は、すべてのサンプルを**「平均スペクトルからの“ずれ”」**として表現し直す操作である。

- **幾何学的意味**: 元のデータ点群（スキャッタープロット）の重心を、グラフの原点 $(0, 0, \ldots, 0)$ に移動させる操作に等しい。
- **なぜ必要か**: PCA や共分散は「平均からのばらつき」を扱う理論であるため、中心化はその必須の前準備となる。

C.2.3 共分散行列の読み方

共分散行列 $\mathbf{S} = \frac{1}{n-1}\mathbf{X}_c^T\mathbf{X}_c$ の各要素には、明確な化学的意味がある。

- **対角成分 (S_{jj})**: 波長 j における吸光度の「ばらつき（分散）」。この値が大きい波長ほど、サンプル間での変化が大きいことを意味する。
- **非対角成分 (S_{jk})**: 波長 j と波長 k の「連動性（共分散）」。
  - **正**: j が増えると k も増える（同じ化学成分由来のピークなど）。
  - **負**: j が増えると k は減る。

スペクトルにおける意味: 化学構造由来のピークは「ブロード（幅広）」であるため、隣接する波長同士は極めて強い正の相関（共線性）を持つ。PCA はこの「共分散のパターン」をまとめて扱うことで、多重共線性の問題を解決する。

C.3　固有分解と PCA とのつながり

C.3.1　固有値・固有ベクトルの直感

行列方程式 $\mathbf{Su} = \lambda\mathbf{u}$ は、以下のことを主張している：

- 「データのばらつきを、方向 $\mathbf{u}$ に沿って見ると、その大きさ（分散）が λ になる」

PCA はこれを繰り返し行う：

1. 一番ばらつきが大きい方向を探す $\rightarrow$ 第一主成分 $\mathbf{p}_1$
2. 次に大きい（かつ $\mathbf{p}_1$ に直交する）方向を探す $\rightarrow$ 第二主成分 $\mathbf{p}_2$

C.3.2　化学的にどう解釈するか

数式上の「固有ベクトル」は、化学者にとっての「スペクトルパターン」に対応する。

- **第一主成分 ($\mathbf{p}_1$)**: データの「全体の濃淡」や「最も支配的な成分の変動」を表すことが多い。
- **第二・第三主成分**: 「より細かい化学構造の違い」や「少量の不純物」に対応することが多い。

ローディングプロットにおいて、固有ベクトルのピーク位置を既知の吸収帯（O-H, C-H 等）と照らし合わせることで、統計的な結果に化学的な意味を与えることができる。

C.4　線形回帰 (OLS) とその限界の明示

C.4.1　行列表現の意味

線形モデル $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ は、構成要素を以下のように整理できる：

- $\mathbf{y}$: 知りたい答え（濃度、オクタン価など）。
- $\mathbf{X}$: 手がかりとなるデータ（スペクトル）。
- $\mathbf{b}$: 各波長の「寄与度」（回帰係数）。ここを知りたい。
- $\mathbf{e}$: 測定誤差やモデル化できなかった「残りかす」。

OLS 解 $\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ は、**「誤差 $||\mathbf{e}||^2$ が最小になるような $\mathbf{b}$ を選びなさい」**という最適化問題の答えである。

C.4.2　HDLSS で何が壊れるか

- 変数の数 d がサンプル数 n より多い場合 ($d \gg n$)、$\mathbf{X}^T\mathbf{X}$ は「ランク落ち」と呼ばれる状態になり、逆行列が存在しない（計算不能になる）。
- これは、スペクトルの「多重共線性」が強すぎて、解が一意に定まらないことの数学的な現れである。

ここから、「だから PCA や PLS による次元削減が必要になる」というケモメトリクスの核心へとつながる。

付録 D　ケモメトリクス手法の理論的定式化

本付録では、本文で議論されたケモメトリクス手法の厳密な数学的定式化と主要な理論的結果を提供する。高次元小標本 (HDLSS) 環境に関連する幾何学的および統計的性質に焦点を当てる。

D.1　主成分分析 (PCA)

D.1.1　幾何学的最適化

PCA は、射影されたデータの分散を最大化する直交基底ベクトル $\{\mathbf{p}_1, \ldots, \mathbf{p}_d\}$ のセットを探す。中心化されたデータ行列を $\mathbf{X}_c \in \mathbb{R}^{n \times d}$ とする。第 1 主成分 $\mathbf{p}_1$ は以下の解である：

$$\mathbf{p}_1 = \arg\max_{\mathbf{u}, ||\mathbf{u}||=1} \mathrm{var}(\mathbf{X}_c \mathbf{u}) = \arg\max_{\mathbf{u}, ||\mathbf{u}||=1} \mathbf{u}^T \mathbf{S}_n \mathbf{u} \tag{17}$$

ここで $\mathbf{S}_n = \frac{1}{n-1} \mathbf{X}_c^T \mathbf{X}_c$ は標本共分散行列である。

D.1.2　双対解釈と計算効率

定理 付録 D.1 (双対固有値関係)**.** 行列 $\mathbf{A}$ の j 番目の非ゼロ固有値を $\lambda_j(\mathbf{A})$ と表記する。このとき：

$$\lambda_j(\mathbf{X}_c^T \mathbf{X}_c) = \lambda_j(\mathbf{X}_c \mathbf{X}_c^T) \tag{18}$$

この双対性により、$d \times d$ の共分散行列の代わりに $n \times n$ のカーネル行列 $\mathbf{K} = \mathbf{X}_c \mathbf{X}_c^T$ を介して PCA を計算することができ、計算量が $O(d^3)$ から $O(n^3)$ に削減される。

D.1.3　HDLSS における不一致性

高次元統計学における重要な結果は、n を固定して $d \to \infty$ としたとき、標準的な PCA は一致性を持たないということである。

定理 付録 D.2 (標本固有値の不一致性)**.** 母固有値が $\lambda_j \sim d^\alpha$ *(*$\alpha \in (0.5, 1]$*)* として振る舞う強スパイク固有値 *(SSE)* モデルの下では、標本固有値 $\hat{\lambda}_j$ はバイアスのある推定量となる：

$$\frac{\hat{\lambda}_j}{\lambda_j} \xrightarrow{P} 1 + \frac{tr(\Sigma_{noise})}{\lambda_j} \neq 1 \tag{19}$$

($d \to \infty$ *のとき)*。

これにより、本文で議論されたノイズ削減 (NR) 法や幾何学的調整の使用が必要となる。

D.2　部分的最小二乗法 (PLS)

D.2.1　最適化問題

PLS は、予測子空間と応答との間の共分散を最大化する方向を探す。

定義 付録 D.1 (PLS 目的関数)**.** j 番目の PLS 重みベクトル $\mathbf{w}_j$ は以下によって求められる：

$$\max_{\mathbf{w}} \mathrm{cov}^2(\mathbf{X}_{j-1}\mathbf{w}, \mathbf{y}_{j-1}) \quad \text{s.t.} \quad ||\mathbf{w}|| = 1 \tag{20}$$

ここで $\mathbf{X}_{j-1}$ と $\mathbf{y}_{j-1}$ は前のステップからデフレートされた行列である。

D.2.2　クリロフ部分空間による解釈

PLS の優雅な理論的性質は、クリロフ部分空間との関係である。

定理 付録 D.3 (PLS とクリロフ部分空間)**.** 最初の A 個の *PLS* 重みベクトルによって張られる部分空間 $\mathcal{W}_A = span\{\mathbf{w}_1, \ldots, \mathbf{w}_A\}$ は、共分散行列 $\mathbf{S} = \mathbf{X}^T\mathbf{X}$ と相互共分散ベクトル $\mathbf{s} = \mathbf{X}^T\mathbf{y}$ によって生成される A 次のクリロフ部分空間と等価である：

$$\mathcal{W}_A = \mathcal{K}_A(\mathbf{X}^T\mathbf{X}, \mathbf{X}^T\mathbf{y}) = span\{\mathbf{X}^T\mathbf{y}, (\mathbf{X}^T\mathbf{X})\mathbf{X}^T\mathbf{y}, \ldots, (\mathbf{X}^T\mathbf{X})^{A-1}\mathbf{X}^T\mathbf{y}\} \tag{21}$$

これは、PLS がなぜ効率的であるかを説明している。PLS は、内部の分散構造 ($\mathbf{X}^T\mathbf{X}$) を考慮しながら、最大相関 ($\mathbf{X}^T\mathbf{y}$) の方向と整列する基底を反復的に構築する。

D.3 主成分回帰 (PCR)

PCR は、$\mathbf{X}$ のみに基づいて部分空間 $\mathbf{P}_A$ を作成する。推定量は以下のように書ける：

$$\hat{\mathbf{b}}_{PCR} = \sum_{j=1}^{A} \frac{1}{\hat{\lambda}_j} \mathbf{p}_j \mathbf{p}_j^T \mathbf{X}^T \mathbf{y} \tag{22}$$

これはスペクトルカットオフ正則化である。固有ベクトル $\mathbf{p}_{A+1}, \ldots, \mathbf{p}_d$ の寄与をゼロにする。すべての成分を連続的に縮小するリッジ回帰とは異なり、PCR は離散的/「ハード」な閾値処理を提供する。

参考文献

[1] Jeongyoun Ahn, James Stephen Marron, Keith M Muller, and Y-Y Chi. High-dimension, low-sample-size data analysis: the geometrical representation of distinct classes. *Annals of statistics*, pages 2886–2908, 2007.

[2] Makoto Aoshima and Kazuyoshi Yata. Statistical inference for high-dimension, low-sample-size data with singular value decomposition. *Japanese Journal of Statistics and Data Science*, 1(1):229–250, 2018.

[3] RJ Barnes, MS Dhanoa, and Susan J Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied spectroscopy*, 43(5):772–777, 1989.

[4] Rasmus Bro. Parafac. tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(1):149–171, 1997.

[5] Hyonho Chun and Sunduz KeleÅ§. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

[6] David L Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1:32, 2000.

[7] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[8] Paul Geladi and Bruce R Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17, 1986.

[9] Paul Geladi, D MacDougall, and H Martens. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy*, 39(3):491–500, 1985.

[10] Peter Hall, JS Marron, and Arik Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444, 2005.

[11] Bruce R Kowalski and CF Bender. Pattern recognition. a powerful approach to interpreting chemical data. *Journal of the American Chemical Society*, 94(16):5632–5639, 1972.

[12] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[13] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[14] Johan Trygg and Svante Wold. Orthogonal projections to latent structures (o-pls). *Journal of chemometrics*, 16(3):119–128, 2002.

[15] Herman Wold. Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, 1:391–420, 1966.

[16] Svante Wold. Chemometrics. *Grant application to the Swedish Natural Science Research Council*, 1972.

[17] Svante Wold et al. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743, 1984.

[18] Kazuyoshi Yata and Makoto Aoshima. Effective pca for high-dimension, low-sample-size data with noise reduction. *Journal of multivariate analysis*, 105(1):193–215, 2012.