高精度ソフトセンサ開発に関する研究

京都大学大学院情報学研究科　藤原幸一

1　はじめに

高炉の安定操業は，鉄鋼プロセスにおける最も重要な課題のひとつであるが，その実現には高炉のダイナミクスを表現できる高精度な数理モデルの構築が欠かせない．特に出銑口溶銑温度は，銑鉄成分制御にとって重要であるが，操業条件を変化させてから溶銑温度が変化するまでに数時間以上の時間遅れがあるため，溶銑温度制御には現在の操業状態から数時間後の溶銑温度を予測することが求められる．しかしながら，高炉は大規模かつ複雑なプロセスであるため，物理化学現象に基づいた第一原理モデルの構築は容易ではない．

化学プロセスや半導体プロセスでは，第一原理モデルではなく，統計的手法に基づいてオンライン測定が困難な製品品質やそれらに大きく影響する変数を推定するソフトセンサと呼ばれる技術が広く使われている．たとえば化学プロセスでは，蒸留塔の製品組成推定などにソフトセンサが利用されている．製薬プロセスにおいても，スペクトルデータから製剤中の有効成分量などを推定できる統計モデルの活用が進んでいる．ソフトセンサ設計には，部分的最小二乗法（Partial Least Squares; PLS）が利用されることが多いが，PLS では最小二乗法で問題となる多重共線性の問題を回避し，少数の潜在変数で高精度の統計モデルを構築できるためである [1, 2]．しかし PLS を用いたとしても，高炉のモデリングは容易ではない．その原因のひとつとして，規模の大きなプロセスである高炉では測定される変数が多いことが挙げられる．出力の推定に寄与しない変数をソフトセンサの入力として多く採用してしまうと，推定精度が低下してしまう．そのため，ソフトセンサ構築には適切に入力変数を選択する必要があるが，この問題は組み合わせ的計算となるため，変数の数が多くなると可能な組み合わせが指数関数的に増加する．

一般に入力変数の数を増加させるにつれ，ソフトセンサのモデル構築用サンプルに対するフィッティング性能は向上する．しかし，出力変数と物理的に関係のない変数まで入力変数として用いると，未知サンプルに対する予測性能は低下する．ソフトセンサ設計では適切な入力変数の組み合わせを選択する必要があるが，しばしば入力変数選択は試行錯誤に頼らざるを得ないため，現場には負担の大きな作業であった．したがって，ソフトセンサの予測精度改善および設計効率化のため，システマティックな入力変数選択手法の開発が望まれる．

これまでに遺伝的アルゴリズム（Genetic Algorithm; GA）を用いた変数選択手法が提案されているが [3]，GA によって変数選択の組み合わせの数を削減できたとしても，計算負荷は小さくない．統計学では，変数選択手法としてステップワイズ [4] や Lasso（Least Absolute Shrinkage and Selection Operator）[5] が知られている．一方，PLS に基づいた変数選択手法として，PLS-Beta や VIP（Variable Influence on Projection），SR（Selectivity Ratio）が提案されている [6, 7]．これらの手法は，候補である変数を線形回帰モデルの入力変数として採用するか個別に評価する．しかし，変数間には相関関係が存在するため，変数を個別に入力変数として選択するのは適切ではない場合がある．Lasso を拡張した手法として，いくつかの変数グループから入力変数として採用する変数グループを選択する手法が提案されている．これを group Lasso と呼ぶ [8]．group Lasso を用いるには事前に変数グループを構築する必要があるが，どのように変数グループを構築してよいかは一般に明らかではなかった．

一方，NC スペクトラルクラスタリング（NCSC）を用いて変数間の相関関係に従って変数を分類して変

数グループを構築し，変数グループごとに入力変数として採用するか判定する変数選択手法が提案されている [9]．これを NCSC 型変数選択（NCSC-Based Variable Sepection; NCSC-VS）と呼ぶ．NCSC-VS は化学プロセスにおけるソフトセンサ設計に適用され，従来の変数選択手法より推定精度の高いソフトセンサを構築できる [9]．しかし，NCSC-VS は調整パラメータが多く調整が困難であるという問題があった．

そこで本研究では，NCSC-VS のパラメータ調整を簡略化するために新たな手法を提案する．提案する手法は，NC 法で得られる入力変数と出力変数との相関関係の類似度に基づいて，入力変数に重みを付けモデル構築を行うもので，NC 変数重み付け（NCSC-Based Variable weighting; NCVW）と呼ぶ．NCVW では変数選択を行わないために，選択のための閾値などの調整パラメータが存在しない．そのため，調整パラメータは NC 法の有するパラメータの 1 つのみであり，ソフトセンサ構築のための手間を大幅に削減できる．

本報告書では，提案法をデータの機密性の高い高炉プロセスではなく，医薬品製造プロセスにおけるソフトセンサ設計に適用し，従来法と比較した結果を報告する．高炉プロセスへの適用結果は，しかるべきデータ開示手続きを経た上で，将来的に論文として公表する．

2　部分的最小二乗法（PLS）

本節では PLS について説明する．PLS では，入力 $\boldsymbol{X} \in \Re^{N \times M}$ と出力 $\boldsymbol{y} \in \Re^{N}$ を次式のように分解する．

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \tag{1}$$

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{b} + \boldsymbol{f} \tag{2}$$

ここで，$\boldsymbol{T} \in \Re^{N \times K}$ は潜在変数 $\boldsymbol{t}_k \in \Re^{N}$ $(k = 1, \cdots, K)$ を並べた行列，$\boldsymbol{P} \in \Re^{M \times K}$ は $\boldsymbol{X}$ のローディング $\boldsymbol{p}_k \in \Re^{M}$ を並べた行列，$\boldsymbol{b} = [b_1, \cdots, b_K]^T \in \Re^{K}$ は $\boldsymbol{y}$ のローディングである．また，K は採用する潜在変数の数であり，$\boldsymbol{E} \in \Re^{N \times M}$ と $\boldsymbol{f} \in \Re^{N}$ は誤差である．

PLS モデルの構築には，NIPALS（Nonlinear Iterative Partial Least Squares）アルゴリズムが用いられる [1]．第 1 番目から第 k 番目までの潜在変数を $\boldsymbol{t}_1, \cdots, \boldsymbol{t}_k$，ローディングを $\boldsymbol{p}_1, \cdots, \boldsymbol{p}_k$，および $b_1, \cdots, b_k$ とする．第 $k+1$ 番目の入力と出力の残差は

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \boldsymbol{t}_k \boldsymbol{p}_k^T \tag{3}$$

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - b_k \boldsymbol{t}_k \tag{4}$$

と書ける．ここで，$\boldsymbol{X}_1 = \boldsymbol{X}$，$\boldsymbol{y}_1 = \boldsymbol{y}$ である．また，$\boldsymbol{t}_k = \boldsymbol{X}_k \boldsymbol{w}_k$ であり，$\boldsymbol{w}_k \in \Re^{M}$ は第 k 番目の重みベクトルである．$\boldsymbol{w}_k$ は $||\boldsymbol{w}_k|| = 1$ の制約下で $\boldsymbol{y}_k$ と $\boldsymbol{t}_k$ の共分散を最大化するように定義される．Lagrange の未定乗数法より

$$G_k = \boldsymbol{y}_k^T \boldsymbol{X}_k \boldsymbol{w}_k - \mu(||\boldsymbol{w}_k||^2 - 1) \tag{5}$$

を最大化すればよい．なお，μ は未定乗数である．$\partial G_k / \partial \boldsymbol{w} = 0$ を解くと

$$\boldsymbol{w}_k = \frac{\boldsymbol{X}_k^T \boldsymbol{y}_k}{||\boldsymbol{X}_k^T \boldsymbol{y}_k||} \tag{6}$$

が得られる．よって第 k 番目のローディング $\boldsymbol{p}_k$ と b_k は

$$\boldsymbol{p}_k = \frac{\boldsymbol{X}_k^T \boldsymbol{t}_k}{\boldsymbol{t}_k^T \boldsymbol{t}_k},\ b_k = \frac{\boldsymbol{y}_k^T \boldsymbol{t}_k}{\boldsymbol{t}_k^T \boldsymbol{t}_k} \tag{7}$$

となる．最終的に採用する潜在変数の数 K に到達するまで上記手順を繰り返し，$\boldsymbol{p}_k$ および b_k を求めればよい．なお，K はクロスバリデーションを用いて決定することができる．

3　従来の変数選択手法

3.1　PLS-Beta

PLS-Beta では PLS モデル Eqs. (1)，(2) を重回帰（multiple linear regression; MLR）モデルに変換し，MLR モデルの回帰係数の大きさに基づいて，PLS モデルの入力変数を選択する [6]．MLR モデルの出力予測値は

$$\hat{\boldsymbol{y}} = \boldsymbol{T}(\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{T}\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_{pls} \tag{8}$$

と書かれる．ここで回帰係数は

$$\boldsymbol{\beta}_{pls} = \boldsymbol{W}(\boldsymbol{P}^T\boldsymbol{W})^{-1}(\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{y} \tag{9}$$

であり，$\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K] \in \Re^{M\times K}$ である．

入力変数は回帰係数 $\boldsymbol{\beta}_{pls}$ の降順に，たとえば

$$\frac{||\boldsymbol{\beta}_{select}||}{||\boldsymbol{\beta}_{pls}||} > \mu \ (0 < \mu \leq 1) \tag{10}$$

となるまで個別に選択する．ここで，$\boldsymbol{\beta}_{select}$ は選択された入力変数に対応する回帰係数ベクトルである．

3.2　VIP

VIP は入力の出力への寄与を表すスコアを用いて，PLS モデルの入力変数を選択する手法である [6]．第 j 番目の入力変数候補のスコア V_j は

$$V_j = \sqrt{M\sum_{k=1}^{K}\left(w_{jk}^2 b_k^2(\boldsymbol{t}_k^T\boldsymbol{t}_k)/||\boldsymbol{w}_k||^2\right)\Big/\sum_{k=1}^{K} b_k^2(\boldsymbol{t}_k^T\boldsymbol{t}_k)} \tag{11}$$

と定義される．ここで w_{jk} は重みベクトル $\boldsymbol{w}_k$ の第 j 番目の要素である．VIP では，たとえば $V_j > \eta(> 0)$ となる変数を入力変数として選択する．

3.3　SR

SR は，PLS モデル Eq. (1) より，入力変数は潜在変数の線形結合と誤差で表現できることに着目した変数選択手法である [7]．SR では，入力変数の分散のうち潜在変数で説明される部分 v_j^{expl} と誤差で説明される部分 v_j^{err} の比 $s_j = v_j^{expl}/v_j^{err}$ をスコアとして定義し，スコア s_j の大きさに従って入力変数を選択する．

入力変数はスコア s_j の降順に，たとえば

$$\frac{||\boldsymbol{s}_{select}||}{||\boldsymbol{s}_{all}||} > \xi \ (0 < \xi \leq 1) \tag{12}$$

となるまで個別に選択する．ここで，$\boldsymbol{s}_{all}$ と $\boldsymbol{s}_{select}$ は，それぞれ全ての変数と選択された入力変数に対応するスコアベクトルである．

3.4 ステップワイズ

ステップワイズは仮説検定に基づいて，MLR モデルの入力変数を選択する方法である．ステップワイズでは，入力変数候補を追加したり取り除いたりしながら，F 検定を用いて追加した入力変数候補の回帰係数の真値が零であるか検定する [4]．ステップワイズでは検定における p 値の閾値 $\bar{p}$ を，調整パラメータとできる．

3.5 Lasso

Lasso は，1 次の正則化の下で 2 乗誤差を最小とする回帰係数 $\boldsymbol{\beta}$ を求める回帰手法である [5]．Lasso の目的関数は

$$\boldsymbol{\beta}_{lasso} = \arg\min_{\boldsymbol{\beta}} \left(||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \right) \tag{13}$$

と定式化される．ただし，$\lambda(>0)$ はパラメータである．Lasso ではいくつかの回帰係数が零となりやすく，この性質より変数を選択する．

3.6 group Lasso

group Lasso は，個別に変数を選択するのではなく，いくつかの変数グループから入力変数グループを選択するように Lasso を拡張した手法である [8]．いま，M 個の変数を J 個のグループに分割し，j 番目の変数グループに属する変数の数を M_j $(M = \sum_{j=1}^{J} M_j)$ とする．$\boldsymbol{X}_j$ と $\boldsymbol{\beta}_j$ は，それぞれ j 番目の変数グループの入力変数行列と回帰係数ベクトルである．group Lasso の目的関数は

$$\boldsymbol{\beta}_{glasso} = \arg\min_{\boldsymbol{\beta}} \left(||\boldsymbol{y} - \sum_{j=1}^{J} \boldsymbol{X}_j \boldsymbol{\beta}_j||_2^2 + \lambda \sum_{j=1}^{J} \sqrt{M_j} ||\boldsymbol{\beta}_j||_2 \right) \tag{14}$$

と定式化される．ここで，$\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_J^T]^T$ であり，λ (>0) はパラメータである．group Lasso を用いて変数を選択するためには，変数グループを予め構築しておく必要がある．

4 NC スペクトラルクラスタリング（NCSC）

本研究では，変数グループ構築に NC スペクトラルクラスタリング（NCSC）[10, 11] を用いる．NCSC は変数間の相関関係を指標としたパターン認識手法である相関識別法（Nearest Correlation Method; NC 法）[12] と，重み付きグラフの分割法であるスペクトラルクラスタリング（SC）[13] を組み合わせた手法であり，変数間の相関関係に基づいてサンプルをクラスタリングできる．

4.1 スペクトラルクラスタリング（SC）

SC は重み付きグラフの分割法であり，類似度を重みとする重み付きグラフのアークをカットすることで，グラフをサブグラフに分割する．SC にはいくつかのアルゴリズムが知られているが，ここでは Max-Min（Mcut）法 [13] について紹介する．

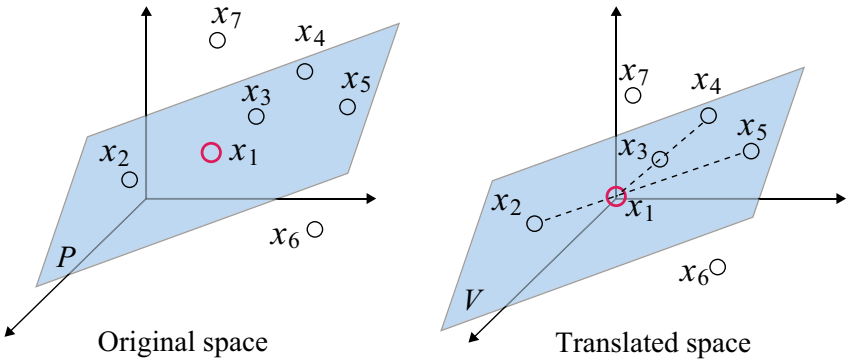

図 1　An example of the procedure of the NC method

いま，重み付きグラフ G とその隣接行列 $\boldsymbol{W}$ が与えられ，G をサブグラフ A と B に分割する問題を考える．まず，サブグラフ間類似度 $\mathrm{cut}(A,B)$ を，サブグラフ間に存在するアークの重みの総和と定義する．すなわち $\mathrm{cut}(A,B) \equiv W(A,B)$ とする．ただし $W(A,B) = \sum_{u\in A, v\in B} W_{u,v}$ である．一方，サブグラフ内類似度を，それぞれのサブグラフ内に存在するアークの重みの総和とする．すなわち，$W(A) \equiv W(A,A)$ である．Mcut 法は $\mathrm{cut}(A,B)$ を最小，かつ $W(A)$ 及び $W(B)$ を最大化するようなサブグラフ A,B を探索する方法である．Mcut 法の目的関数 J は次式で定義される．

$$\min\ J = \frac{\mathrm{cut}(A,B)}{W(A)} + \frac{\mathrm{cut}(A,B)}{W(B)} \tag{15}$$

この最小化問題は行列の固有値問題に帰着される [13].

4.2　相関識別（NC）法

NC 法は，いくつかの異なった相関関係を有するサンプルから構成されているデータにおいて，クエリと相関関係が類似のサンプルを検出する手法である [12].

NC 法のコンセプトを述べる．図 1 (左) のアフィン部分空間 P が変数間の相関関係を表しており，P 上のサンプルはすべて同一の相関関係に従っているものとする．すなわち，$\boldsymbol{x}_1,\cdots,\boldsymbol{x}_5$ は同一の相関関係に従うサンプルであるが，$\boldsymbol{x}_6$，$\boldsymbol{x}_7$ は異なった相関関係を有している．

いま，サンプル $\boldsymbol{x}_1$ と類似の相関関係を有するサンプルを検出したいとする．つまり，$\boldsymbol{x}_2,\cdots,\boldsymbol{x}_4$ が検出できればよい．NC 法ではまず，$\boldsymbol{x}_1$ が原点となるように空間全体を平行移動させる．これは全サンプル $\boldsymbol{x}_i (i=1,\cdots,7)$ から $\boldsymbol{x}_1$ を引けばよい．この操作によってアフィン部分空間 P は原点を含むことになるため線形部分空間の定義を満たす．これを V とする．

次に，図 1 (右) に示すように，任意のサンプルと原点を結ぶ直線を引く．いま，この直線上で別のサンプルが発見できたとする． この例では，$\boldsymbol{x}_2$ - $\boldsymbol{x}_5$ 及び $\boldsymbol{x}_3$ - $\boldsymbol{x}_4$ がこのような関係を満たしている．このとき，これらのサンプルのペアの相関係数の絶対値は 1 である．一方，V の要素ではない $\boldsymbol{x}_6$，$\boldsymbol{x}_7$ の相関係数の絶対値は 1 未満である．それゆえ，相関係数が ± 1 であるペアのサンプルは，同一の相関関係を有していると判定できる．

実際は相関係数が厳密に ± 1 になるペアは存在しないため，閾値 γ $(0<\gamma\leq 1)$ を用いて，同一の相関関係を有しているかを判定する．すなわち，相関係数 $C_{i,j}$ について $|C_{i,j}| > \gamma$ 以上となるペア $\boldsymbol{x}_i$ - $\boldsymbol{x}_j$ を，クエリと類似の相関関係を有していると判定する．

4.3 NC スペクトラルクラスタリング（NCSC）

NC 法を用いることで，類似の相関関係に従うサンプルのペアを検出できる．よって，NC 法の結果から変数間の相関関係を指標として SC のための類似度行列を構成する．いま，サンプル $\boldsymbol{x}_n \in \Re^M$ $(n = 1, \cdots, N)$ がデータベースに保存されているとする．NCSC のアルゴリズムを Algorithm 1 に示す．

5 NCSC 型変数選択（NCSC-VS）

従来の変数選択手法は，ソフトセンサの入力変数として採用すべきか変数を個別に評価する．しかし，一般的に入出力間だけではなく入力変数間にも相関関係が存在するため，ある入力変数が変化すると，出力変数のみならず他の入力変数も同時に変化してしまう．つまり，個別に変数を選択するのは必ずしも適切ではないため，入力変数を複数同時に選択する必要がある．

そこで，NCSC-VS では，NSCS を用いて変数間の相関関係に従って変数をいくつかの変数グループに分類し，変数グループごとに線形回帰モデルの入力変数として採用するか評価する．まず，入力変数候補を NCSC を用いて J 個の変数グループ $\boldsymbol{v}_j = \{x_m \mid m \subset \mathcal{V}_j\}$ $(j = 1, \cdots, J)$ に分類する．$\mathcal{V}_j$ は j 番目のグループに属する変数のインデックスの集合であり，$\mathcal{V} = \cup_j \mathcal{V}_j$ である．NCSC はサンプルをクラスタリングする手法であるから，NCSC-VS では入力変数候補をクラスタリングするために，データ行列 $\boldsymbol{X}$ の転置 $\boldsymbol{X}^T$ を NCSC の入力とする．

次に，第 j 番目のクラス $\boldsymbol{v}_j$ の要素の変数から構築したデータ行列 $\boldsymbol{X}_j$ より，第 j 番目の PLS モデル $f_{j,k}$ を構築し，$f_{j,k}$ が出力予測に寄与しているかを評価する．このとき，構築する PLS モデルの潜在変数の数を k とする．ここで寄与率を

$$C_{j,k} = 1 - \frac{||\hat{\boldsymbol{y}}_{j,j}||^2}{||\boldsymbol{y}||^2} \tag{16}$$

と定義する．ここで，$\hat{\boldsymbol{y}}_{j,k}$ は $f_{j,k}$ による予測値である．最終的に $C_{j,k}$ の降順に D $(\leq J)$ 個の変数グループを選択し，選択された変数グループの要素を入力変数としてモデルを構築する．この方法を NCSC 型変数選択

Algorithm 1 Nearest correlation spectral clustering (NCSC)

Set $\boldsymbol{S} \in \Re^{N \times N} \leftarrow \boldsymbol{O}_{N,N}$, γ $(0 < \gamma \leq 1)$ and $L = 1$.
for $L = 1$ to N **do**
  Set $\boldsymbol{S}_L \in \Re^{N \times N} \leftarrow \boldsymbol{O}_{N,N}$.
  for all $n = 1, 2, \cdots, N$ $(n \neq L)$ **do**
    $\boldsymbol{x}'_n = \boldsymbol{x}_n - \boldsymbol{x}_L$.
  end for
  for all k, l $(k \neq l)$ such that $|C'_{k,l}| \geq \gamma$ **do**
    $(\boldsymbol{S}_L)_{k,l} = (\boldsymbol{S}_L)_{l,k} = 1$.
  end for
  $\boldsymbol{S} = \boldsymbol{S} + \boldsymbol{S}_L$.
end for
Partition $\boldsymbol{S}$ by SC.

（NCSC-VS）と呼ぶ．提案する NCSC-VS では，NC 法の閾値 γ，NCSC にて構築する変数グループの数 J，i 番目の変数グループの PLS モデル $f_{j,k}$ の潜在変数の数 k および採用する変数グループの数 D の 4 つが調整パラメータとなる．

6　NCSC 型変数重み付け（NCVW）

NCSC-VS は，変数間の相関関係を考慮して適切にソフトセンサの入力変数を選択し，ソフトセンサの性能を改善することができるが，調整パラメータが 4 つあるため，その最適な選択には手間がかかることが問題であった．そこで本研究では，入力変数を選択するのではなく，NC 法を用いて入力変数に適切な重み付けを行うことで，ソフトセンサの性能を改善できる手法である NCVW を提案する．

NCSC では入力変数間の相関関係を NC 法を用いて判別し，重み行列を構築した．NCVW では入力変数に加えて出力変数を加えて NC 法を実施することで，入出力変数間の相関関係に基づいて類似度を計算する．そして，計算したそれぞれの入力変数を重みベクトルとみなし入力変数を重み付けした変数を入力として，PLS によってソフトセンサを構築する．NCVW では入力変数間の相関関係のみならず，出力変数と相関を有する入力変数を重要視してモデリングを行うことで，ソフトセンサの性能を改善できる．

提案する NCVW では，入力変数と出力変数をまとめて新たに NC 法の入力とする．すなわち M 個の入力変数が存在するとき n 番目のサンプルを

$$\boldsymbol{x}'_n = [x_n^{[1]}, \cdots, x_n^{[M]}, y_n]^T \tag{17}$$

として NC 法を適用し，類似度行列 $\boldsymbol{S}$ を構築する．そして，$\boldsymbol{S}$ の $M+1$ 列目 $\boldsymbol{s}^{[M+1]}$ の第 1〜第 M 成分を重みベクトル $\boldsymbol{w} = [w^{[1]}, \cdots, w^{[M]}$ として抽出する．最終的に新たな入力変数を

$$z_n = [w^{[1]}x^{[1]}, \cdots, w^{[M]}x^{[M]}]^T \tag{18}$$

として．PLS にてソフトセンサ構築を行う．

7　医薬品製造プロセスにおける検量線設計への適用

本ケーススタディでは，医薬品製造プロセスにおける検量線設計を通じて，提案する NCVW と従来法との比較を行った．対象プロセスは第一三共株式会社の製剤混合プロセスであり，製剤の有効成分（active pharmaceutical ingredient; API）含有量を推定する検量線を設計するために，入力波長選択を行った [15].

7.1　解析対象データ

対象とする製剤混合プロセスでは 6 種類の成分を混合している．実験によって異なる API 含有量の製剤を調製しその時の NIR スペクトル（800−2500 nm・2203 波長）を測定した．得られたモデル構築用データ，モデル検証用データはそれぞれ 576 サンプル，および 20 サンプルである．前処理として，1 次の Savitzky-Golay フィルタ [16] を用いてスペクトルデータを平滑化した．

7.2　波長選択およびモデル構築

波長選択のベンチマークとして全ての波長を用いて PLS モデルを構築した．これを PLS-All と呼ぶことにする．従来法である BLS-Beta，VIP，ステップワイズ，Lasso および group Lasso を用いて波長選択を

表 1　Estimation results, the selected parameters and the computational load in eight methods

	RMSE	r	#Wavelength	#LV	Parameters
PLS-All	1.23	0.91	2203	37	−
PLS-Beta	1.06	0.90	928	36	$\mu = 0.75$
VIP	1.01	0.91	1133	19	$\eta = 0.8$
SR	1.19	0.91	612	20	$\xi = 0.75$
Lasso	0.98	0.93	1138	39	$\lambda = 0.2$
group Lasso	1.04	0.95	1457	20	$J = 7$, $D = 2$
stepwise	1.42	0.85	561	24	$\bar{p} = 0.15$
NCSC-VS	0.77	0.96	843	25	$\gamma = 0.99$, $J = 6$, $D = 2$
NCVW	0.74	0.96	2203	15	$\gamma = 0.99$

行った．

PLS-Beta，VIP と SR の調整パラメータ μ，η，ξ は，それぞれ $\mu = \{0.70,\ 0.75,\ 0.80,\ 0.85,\ 0.90,\ 0.95\}$，$\eta = \{0.6,\ 0.7,\ 0.8,\ 0.9,\ 1.0,\ 1.1\}$，$\xi = \{0.70,\ 0.75,\ 0.80,\ 0.85,\ 0.90,\ 0.95\}$ とした．ステップワイズ法における p 値の閾値は，$\bar{p} = \{0.005,\ 0.05,\ 0.08,\ 0.1,\ 0.12,\ 0.15\}$ である．Lasso のパラメータは $\lambda = \{0.1,\ 0.2,\ 0.4,\ 0.5,\ 0.8,\ 1.0\}$ であり，回帰係数が 10^{-4} 未満である波長を除去して残りを入力波長とした．ここで，分割数は $J = \{5,\ 6,\ 7,\ 8,\ 9,\ 10\}$ であり，パラメータは $\lambda = \{20,\ 25\}$ である．group Lasso では，解析の前に全スペクトルを同じ長さのいくつかの領域に分割し，波長グループとした [17]．ここで，分割数は $J = \{5,\ 6,\ 7,\ 8,\ 9,\ 10\}$ であり，パラメータは $\lambda = \{20,\ 25\}$ である．

NCSC-VS では，まず NCSC を用いて波長グループを構築した．NC 法のパラメータは $\gamma = 0.99$，構築する波長グループの数は $J = \{5,\ 6,\ 7,\ 8,\ 9,\ 10\}$ である．NCSC-VS における採用波長グループ数は $D = \{2,\ 3\}$ であり，このとき波長グループの PLS モデルの潜在変数の数は $k = 10$ とした．一方，提案する NCVW の調整パラメータは，NC 法のパラメータ γ のみであり，NCSC-VS と同様に $\gamma = 0.99$ とした．これらのパラメータは試行錯誤によって決定した．

それぞれのパラメータによって選択された入力波長を用いて，PLS によって検量線を構築した．このとき，最適な潜在変数の数はクロスバリデーションにて決定し，それぞれの調整パラメータごとに望ましいパラメータを選択した．モデル検証用サンプルを用いて構築した検量線の推定性能を検証した．9 種類の手法における最終的な推定結果と，そのときの望ましいパラメータを表 1 に示した．表中，#LV は PLS の潜在変数の数であり，r は測定値と推定値との相関係数を表す．

この結果より，ステップワイズの推定性能は，どのパラメータを選択しても PLS-All よりも低下している．PLS-Beta，VIP，SR，Lasso，group Lasso はそれぞれ PLS-All よりも推定性能を改善しており，変数選択の効果が表れているといえる．NCSC-VS と提案する NCVW では，これらの手法よりもさらに推定性能を改善しており，両者ともに，PLS-All と比較して RMSE が約 40% 改善している．NCVW は調整パラメータが 1 つのみであるにも関わらず，調整パラメータを 4 つ有する NCSC-VS と同程度の性能を達成しており，NCVW によってより効率的に高精度のソフトセンサが構築できることが分かる．

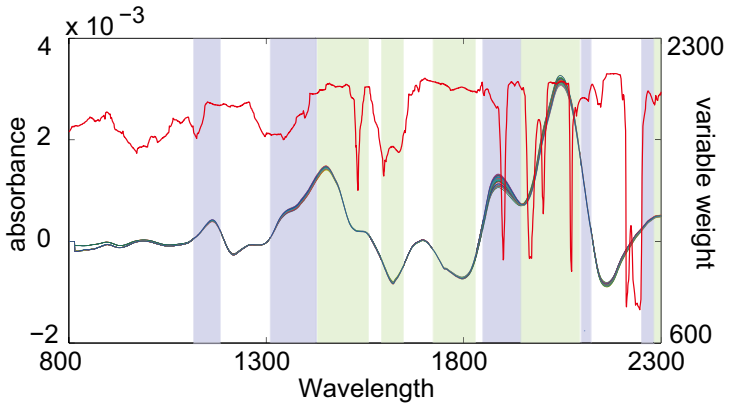

図 2　An example of the procedure of the NC method

7.3　考察

ケーススタディの結果を検証する．図 2 に NCSC-VS で選択された波長と，NCVW にて計算された変数重みを示す．ここで，図中の帯は NCSC-VS にて選択された波長グループであり，カラーは NCSC にて分割された波長グループである．赤線が NCVW での変数重みを示している．この図より，NCSC によって構築された波長グループは，いくつかの連続した波長域からなっており，NCSC-VS にて選択された波長は特定のピークを含んでいることがわかる．一方，NCVW では，大きな重みが与えられている波長ではおおよそピークを含んでいるが，いくつかのピークでは重みが小さくなっていることが分かる．これらの結果は，化合物の情報はスペクトルの特定のピークにより多く含まれるという物理化学的な事実と矛盾しない．すなわち，いくつかのピークは確かに API 推定に重要な情報を含んでいるが，一方で推定に不必要なピークも存在しており，NCVW による変数重み付けの結果は推定に不必要であったピークの存在を示唆していると考えられる．

したがって，NCSC-VS および提案する NCVW は，検量線構築に意味のある波長を選択，または重み付けしていると結論づけられる．

8　結言

本研究では，ソフトセンサ構築を対象として，NC 入力変数重み付け（NCVW）を提案した．さらに高炉プロセス同様に変数が非常に多い医薬品製造プロセスの実データを用いたケーススタディを通じて，その有効性を示した．提案法は,，NC 法で得られる入力変数と出力変数との相関関係の類似度に基づいて，入力変数に重みを付けモデル構築を行うもので，入力選択のための閾値などの調整パラメータが存在しない．そのため，調整パラメータは NC 法の有するパラメータの 1 つのみであり，ソフトセンサ構築のための手間を大幅に削減でき，ソフトセンサの効率的な設計に貢献できる．

謝辞

本研究は公益財団法人 JFE21 世紀財団の助成を受けて行われた．ここに謝意を表します．

参考文献

[1] S. Wold *et al.*: PLS-Regression: a Basic Tool of Chemometrics, *Chemom. Intell. Lab. Syst.*, **58**, 109/130 (2001)

[2] M. Kano, M. Ogawa: The State of the Art in Chemical Process Control in Japan: Good Practice and Questionnaire Survey, *J. Process Control*, [bf 20, 969/982 (2010)

[3] M. Arakawa, Y. Yamashita and K. Funatsu: Genetic Algorithm-based Wavelength Selection Method for Spectral Calibration, *J. Chemometrics*, **25**, 10/19 (2010)

[4] R. R. Hocking: The Analysis and Selection of Variables in Linear Regression, *Biometrics*, **32**, 1/49 (1976)

[5] R. Tibshirani: Regression Shrinkage and Selection via the Lasso, *J. R. Stat. Soc. Series B Stat. Methodol.*, **58**, 267/288, (1996)

[6] S. Wold *et al.*: 3D QSAR in Drug Design; Theory, Methods, and Applications, *ESCOM*, Leiden, Holland (1993)

[7] T, Rajalahti *et al.*: Biomarker Discovery in Mass Spectral Profiles by Means of Selectivity Ratio Plot, *Chemom. Intell. Lab. Syst.*. **95**, 35/48 (2009)

[8] M. Yuan and Y. Lin: Model Selection and Estimation in Regression with Grouped Variables, *J. R. Stat. Soc. Series B Stat. Methodol.*, **68**, 49/67 (2006)

[9] K. Fujiwara, H. Sawada and M. Kano: Input Variable Selection for PLS Modeling Using Nearest Correlation Spectral Clustering, *Chemom. Intell. Lab. Syst.*, **118**, 109/119 (2012)

[10] K. Fujiwara, M. Kano and S. Hasebe: Development of Correlation-based Clustering Method and Its Application to Software Sensing, *Chemom. Intell. Lab. Syst.*, **44**, 130/138 (2010)

[11] K. Fujiwara, M. Kano and S. Hasebe: Correlation-based Spectral Clustering for Flexible Process Monitoring, *J. Process Control*, **21**, 1438/1448 (2011)

[12] K. Fujiwara, M. Kano and S. Hasebe: Development of Correlation-Based Pattern Recognition Algorithm and Adaptive soft-sensor Design, *Control Eng. Pract.*, **20**, 371/378 (2012)

[13] C. H. Q. Ding *et al*: A Min-Max Cut Algorithm for Graph Partitioning and Data Clustering, *IEEE Int'l Conf. Data Min.*, San Jose (2001)

[14] A. N. Ng, M. I. Jordan and Y. Weiss: On Spectral Clustering: Analysis and an Algorithm, *NIPS*, Vancouver (2001)

[15] S. Kim *et al.*: Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection, *Int. J. Pharm.*, **421**, 269/274 (2011)

[16] A. Savitzky and M. J. E. Golay: Smoothing and Differentiation of Data by Simplified Least Squares Procedures, *Anal. Chem.*, **36**, 627/1639 (1964)

[17] L. Norgaard, *et al.*: Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy, *Appl. Spectrosc.*, **54**, 413/419 (2000)