統計的ソフトセンサ構築支援ソフトウェア

SoftsensorBuilder 1.1 の取扱説明書

京都大学　プロセスシステム工学研究室
長谷部伸治，金尚弘
東京大学　船津研究室
船津公人，金子弘昌
京都大学　ヒューマンシステム論分野
加納学，藤原幸一

2016 年 3 月 16 日

目次

1　開発の経緯

ソフトセンサはリアルタイムに測定することが困難な変数を，リアルタイムに測定できる変数を利用して推定するためのものであり，化学，製薬，鉄鋼，半導体など幅広い産業で利用されてきた．また，プロセス中で測定できるデータの量は増加しており，膨大なデータを有効活用したいという要望も増大している．しかし，ソフトセンサの利用に興味があっても実装にまで至らないという事例が散見される．そこで，ソフトセンサの更なる普及を主な目的とし，基本的な統計的ソフトセンサ構築手法を含むソフトウェアを開発した．

2　利用規約

2.1　免責事項

本ソフトウェアの利用によって，利用者が損害を受けたとしても，開発者（京都大学長谷部研究室，東京大学船津研究室，京都大学加納研究室）は責任を負わない．

2.2　権利の帰属先，再配布，編集

開発者は本ソフトウェアの著作権を放棄しない．本ソフトウェアの再配布を希望する際は，事前に開発者に申請しなければならない．本ソフトウェアの編集は可能とするが，編集後のソフトウェアの再配布にも開発者の許可が必要である．

2.3　引用義務

本ソフトウェアを利用した成果を社内外で発表する際には，下記の情報を利用して本ソフトウェアを引用することを，利用者の義務とする．

引用情報（日本語）

京都大学長谷部研究室, 東京大学船津研究室, 京都大学加納研究室,「SoftsensorBuilder Version 1.1」,（2015）

引用情報（英語）

Process Systems Engineering Laboratory, Kyoto University, Funatsu Laboratory, the University of Tokyo and Human Systems Laboratory, Kyoto University, “SoftsensorBuilder, Version 1.1”, (2016)

2.4　アンケート回答義務

利用者は毎年１回実施予定のアンケートに回答しなければならない．秘密保持上の問題などについては，別途開発者と利用者で協議する．

2.5　サポート

メールでのサポートを基本とする．問合わせメールアドレスは以下の通りである．

softsensorbuilder@cheme.kyoto-u.ac.jp

3　配布ファイルの内容

配布しているファイルの内容を表 1 に示す．表 1 中の灰色欄にある文字はフォルダ名であり，各フォルダまたはファイルは左側にあるフォルダに含まれている．MATLAB フォルダ中のファイルを利用するには MATLAB®が必要である．exe フォルダ中のファイルは MATLAB®が無くても利用できる．ただし，MCR_R2014a_win32_installer.exe（32bit パソコンを利用する場合）もしくは MCR_R2014b_win64_installer.exe（64bit パソコンを利用する場合）を利用して，事前に MATLAB Runtime をインストールし，パソコンを再起動する必要がある．上記 2 ファイルのサイズは各々500 MB 程度であり，http://jp.mathworks.com/products/compiler/mcr/ にてダウンロードする必要がある場合もある．図 1 は MATLAB Runtime のダウンロード画面である．稀に，異なるバージョンの MATLAB Runtime が必要となる事例がある．

表 1　配布ファイルの内容

Softsensor Builder_ Ver_2	MATLAB	+SoftsensorBuilder Function	関数ファイル群
		SoftsensorBuilder_Ver_2.fig SoftsensorBuilder_Ver_2.m	
	exe	32 bit	SoftsensorBuilder_Ver_2_win32.exe MCR_R2014a_win32_installer.exe（オプション） splash.png（ソフト起動時画面）
		64 bit	SoftsensorBuilder_Ver_2_win64.exe MCR_R2014b_win64_installer.exe（オプション） splash.png（ソフト起動時画面）
	sample_data.xls, sample_data_DATE.xlsx（デモ用データ） SoftsensorBuilder_取扱説明書_Ver_2.pdf（本書）		

図 1　MATLAB Runtime のダウンロード画面

4　動作環境

表 1 のファイルの動作は下記の環境で確かめた．また，MATLAB Runtime のインストールには約 1.4 GB のディスク容量が必要である．

表 2　動作確認を実施した環境

PC 番号	PC の情報	動作確認をしたファイル
1	OS：Windows 7 Professional 32bit CPU：Intel® Core™ i3-3227U（1.9 GHz） メモリ：4 GB	MTLAB フォルダ内のファイルと SoftsensorBuilder_Ver_2_win32.exe
2	OS：Windows 7 Professional 64bit CPU：Intel® Core™ i7-4510U（2 GHz） メモリ：8 GB	MTLAB フォルダ内のファイルと SoftsensorBuilder_Ver_2_win64.exe

5　事前準備

プロセスデータを xls ファイルに保存しておく．データ保存時の注意事項は下記の通りである．

・xls ファイル中のシート数は 1 とする．

・シートの行番号とサンプル取得時刻，列番号と変数の種類を 1 対 1 対応させる．ただし，先頭行のいくつかは変数名などのテキストデータでも良い．

・シートの 1 列目に出力変数，2 列目以降に入力変数のデータを配置する．
・古いデータから順に小さい行番号（サンプル番号）を対応させる．
・データに欠損値が含まれていてもよいが，欠損値のあるサンプルは無視される．ただし，ソフトウェア中でダイナミックモデルの入力変数行列を作成する時には，出力変数に欠損があるサンプルも利用される．
・ダイナミックモデルを構築する際は入力変数の測定間隔は一定でなければならない．

また，xls ファイルは，exe ファイルを利用する場合は exe ファイルと同一のフォルダに，MATLAB ファイルを利用する場合は MATLAB フォルダもしくは MATLAB のパスが通っているフォルダに保存しておかなければならない．

6　ツールの起動

以下の手順により，図 2 のような GUI が現れる．GUI の表示に 10 秒程度かかる．

6.1　MATLAB フォルダを利用する場合

MATLAB®のパスに MATLAB フォルダを含め，SoftsensorBuilder_Ver_2.m を実行する．関数名の競合に注意する必要がある．競合が起こる場合は，パス設定を変更する．

6.2　exe フォルダを利用する場合

32 bit のシステムを利用する場合には，まず，MCR_R2014a_win32_installer.exe を実行し，MATLAB Runtime をインストールする．その後，SoftsensorBuilder_win32.exe を実行する．64 bit のシステムを利用する場合には，64 bit フォルダにあるファイルで同様の作業を実施すればよい．

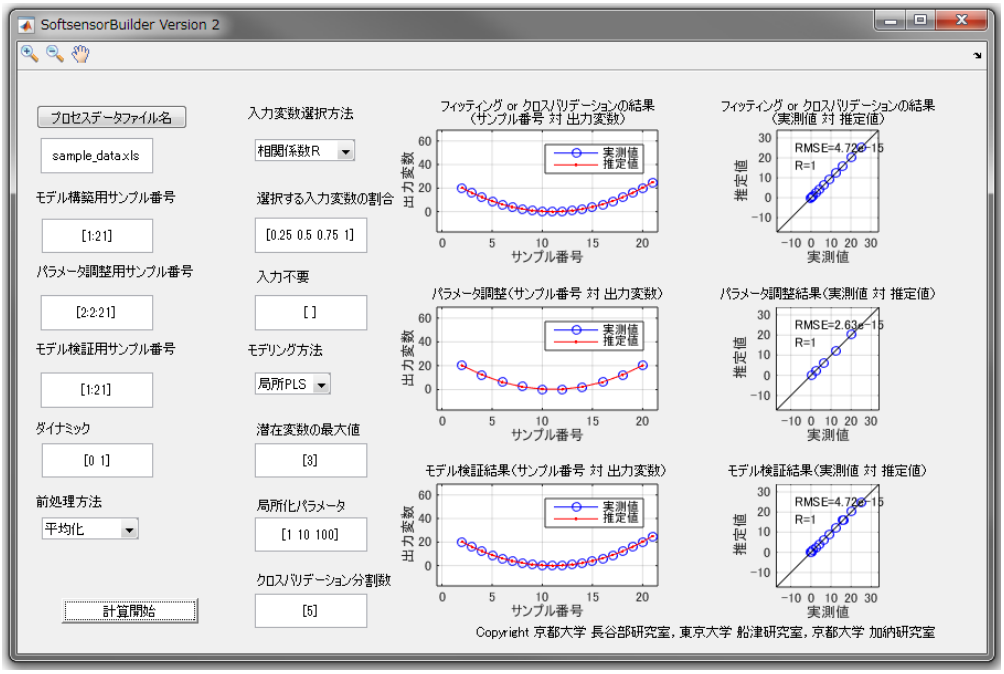

図 2　ソフトセンサ構築プログラムの GUI

7　ソフトセンサの構築

7.1　ソフトセンサ構築手順の概要

GUI 左側にソフトセンサ構築に必要な情報を入力し，計算開始ボタンをクリックすればソフトセンサを構築できる．そうすれば，ユーザーが入力したパラメータの組み合わせが個別に評価される．そして，最も評価の高いパラメータの組み合わせを利用した際の出力推定結果が，GUI 右側のグラフに表示される．また，それぞれのパラメータの組み合わせを利用した場合の推定結果をまとめた xlsx ファイルが生成される．

以下では，ユーザーが入力する項目について説明する．ただし，ベクトルの定義方法は MATLAB と同様である．入力した条件によって計算負荷が変化するので，利用の初期段階では，サンプル数やパラメータの候補数を少ない条件下で予備検討を実施することを推奨する．

7.2　ユーザー入力項目

7.2.1　プロセスデータファイル名

5 節で準備した xls ファイル名を入力する．拡張子もファイル名に含める．もしくは，プロセスデータファイル名というボタンを押し，利用するファイルを選択する．

7.2.2　モデル構築用サンプル番号

モデル構築用のサンプル番号を要素に持つベクトルを入力する．

7.2.3　パラメータ調整用サンプル番号

パラメータ調整用のサンプル番号を要素に持つベクトルを入力する．空配列でも良い．

7.2.4　モデル検証用サンプル番号

モデル検証用のサンプル番号を要素に持つベクトルを入力する．空配列でも良い．

7.2.5　ダイナミック

出力変数の推定に利用する入力変数の測定時間に対応するベクトルを入力する．例えば，ある時刻の出力変数を予測するために，その測定時刻と 2 時刻前の入力変数の測定値を利用する場合，[0 2]と入力すればよい．全ての入力変数に対して同じ時刻が設定される．ダイナミックモデル構築に利用する入力変数と出力変数の作成方法は付録 B にある．

7.2.6　前処理方法

プルダウンリストから前処理方法（平均化，標準化）を選択する．平均化とは全ての変数の平均をゼロとする操作である．標準化は平均化後に，各変数の標準偏差を 1 とする操作である．ただし，標準偏差が 0 の変数の値は全て 0 にする．

7.2.7　入力変数選択方法

プルダウンリストから入力変数選択方法（全て選択，GAPLS，LASSO，VIP，PLS 回帰係数，相関係数 R）を選択する．

7.2.8　入力変数パラメータ 1, 2

7.2.7 節で選択した入力変数選択方法に対応するパラメータを要素に持つベクトルもしくはスカラーを入力する．入力変数選択方法とパラメータの関係は表 3 の通りである．また，パラメータに対する条件を表 4 に示す．世代数と個体数を大きくすると解が良くなるが，計算時間が長くなる．選択する変数の割合と潜在変数の最大値を大きく（小さく）するとオーバー（アンダー）フィッティングの可能性が高くなる．

選択する入力変数の割合にベクトルを入力した場合，ベクトルの各要素の値を利用した時の入力変数選択結果が計算される．VIP もしくは PLS 回帰係数を利用する場合は，潜在変数の数が 1 から最大値までの場合それぞれに対して入力変数選択結果が計算される．

表 3　入力変数選択方法とパラメータの関係

方法	パラメータ	
	1	2
全て選択	なし	なし
GAPLS	世代数	個体数
VIP	選択する入力変数の割合	潜在変数の最大値
LASSO	選択する入力変数の割合	なし
PLS 回帰係数	選択する入力変数の割合	潜在変数の最大値
相関係数 R	選択する入力変数の割合	なし

表 4　入力変数パラメータに対する条件

パラメータ	条件	推奨値
世代数	自然数のスカラー	100
個体数	自然数のスカラー	100
選択する入力変数の割合	0 から 1 の実数	[0.1:0.1:1]
潜在変数の最大値	モデル構築用入力変数行列のランク以下の自然数のスカラー	1-15

7.2.9　モデリング方法

プルダウンリストからモデリング方法（SVR，PLS，局所 PLS）を選択する．

7.2.10　モデルパラメータ 1-4

7.2.9 節で選択したモデリング方法に対応するパラメータを要素に持つベクトルを入力する．モデリング方法とパラメータの関係を表 5 に示す．また，パラメータに対する条件を表 6 に示す．誤差項の重みを大きく（小さく）するとオーバー（アンダー）フィッティングの可能性が高くなる．許容誤差以下の誤差（出力変数のスケーリング後の誤差であることに注意）は 0 として扱われる．局所化パラメータとクロスバリデーション分割数を大きく（小さく）するとオーバー（アンダー）フィッティングの可能性が高くなる．

誤差項の重み，許容誤差，局所化パラメータにベクトルを入力した場合，ベクトルの各要素の値を利用した時のモデリング結果が計算される．PLS もしくは局所 PLS を利用する場合は，潜在変数の数が 1 から最大値までの場合それぞれに対してモデリング結果が計算される．

表 5　モデリング方法とパラメータの関係

方法	パラメータ		
	1	2	3
SVR	誤差項の重み	許容誤差	クロスバリデーション分割数
PLS	潜在変数の最大値	なし	クロスバリデーション分割数
局所 PLS	潜在変数の最大値	局所化パラメータ	クロスバリデーション分割数

表 6　モデルパラメータに対する条件と推奨値

パラメータ	制約条件	推奨値
誤差項の重み	正の実数	2.^[1:1:15]
許容誤差	正の実数	2.^ [-15:1:0]
潜在変数の最大値	モデル構築用入力変数行列のランク以下の自然数のスカラー	5〜15
局所化パラメータ	正の実数	10.^[-2:0.3:1]
クロスバリデーション分割数	2 以上，モデル構築用サンプル数 N 以下の整数のスカラー	5〜N

8　結果の見方

8.1　GUI 上に表示される結果

計算中には，計算の進捗状況を示すダイアログボックスが表示される．計算途中にエラーが生じた場合には，エラー内容を示すダイアログボックスが表示される．ただし，開発者が想定していないエラーが生じた場合には，ダイアログボックスが表示されること無く計算が中止される．計算が中止されたかどうかは Windows タスクマネージャーで確認されたい．

エラー無くソフトセンサ構築が終了すると，GUI 右側のグラフに推定結果が表示される．パラメータ調整用サンプルの有無によって，結果の意味が変わることに注意が必要である．パラメータ調整用サンプル番号が空配列の場合，モデル構築用データに対するクロスバリデーション誤差が最小になるパラメータを利用した時のクロスバリデーション結果が上段に表示され，中段には何も表示されない．パラメータ調整用サンプル番号が空配列でないときは，パラメータ調整用サンプルに対する推定誤差が最小になるパラメータを利用した時のモデル構築用データに対するフィッティング結果が上段に表示され，パラメータ調整用サンプルに対する推定結果が中段に表示される．下段には，モデル検証用サンプルに対する推定結果が表示される．ただし，モデル検証用サンプル番号が空配列の場合には，何も表示されない．

GUI 右側のグラフの最大値と最小値は GUI 左上の虫めがねアイコンを利用することで変更できる．

8.2　エクセルファイル内の結果

GUI 上には，最も評価の高いパラメータの組み合わせを利用した際の推定結果のみしか表示されないが，xlsx ファイルにはより詳細な情報が保存される．xlsx ファイルの内容は表 7 の通りである．例えば，図 2 のように設定し，sample_data.xls を利用して計算を実行すると，sample_data_DATE.xlsx が得られる．ただし，DATE の部分には計算開始時刻が入る．

表 7 結果が保存される xlsx ファイルの内容

ファイル名	シート名	内容
“FILENAME”_“DATE”.xlsx	setting	GUI で設定した項目
	dynamic data	時間遅れを考慮した入力変数行列
	optimum parameter	最適なパラメータの組み合わせ
	Y_est	各パラメータ組に対する出力の推定値
	input variable	各入力変数パラメータに対する変数選択結果
	reg_coef	前処理前基準の回帰係数（PLS 利用時のみ）
	RMSE	各パラメータ組に対する出力の RMSE

付録A　統計的手法の詳細

A.1　　記号の定義

本節で利用する記号の定義を表 A1 に示す．ただし，データに欠損値が無いことを前提としている．

表 A1 記号の定義

記号	意味
$C \in \Re$	SVR における誤差項の重み
$d_n \in \Re$	第 n サンプルとクエリ $\boldsymbol{x}_q$ の距離
$\bar{d} \in \Re$	d_n（$n = 1, 2, \ldots, N$）の平均値
$K \in \Re$	PLS / 局所 PLS で利用される潜在変数の数
$N \in \Re$	サンプル数
$M \in \Re$	入力変数の数
$\boldsymbol{P} = [\boldsymbol{p}_1 \quad \boldsymbol{p}_2 \quad \cdots \quad \boldsymbol{p}_K] \in \Re^{M \times K}$	PLS / 局所 PLS の入力変数再構築行列
$\boldsymbol{p}_k = [p_{1k} \quad p_{2k} \quad \cdots \quad p_{Mk}]^{\mathrm{T}} \in \Re^M$	入力変数再構築行列 $\boldsymbol{P}$ の第 k 列
$\boldsymbol{q} = [q_1 \quad q_2 \quad \cdots \quad q_K]^{\mathrm{T}} \in \Re^K$	PLS / 局所 PLS の潜在変数に対する回帰係数
$R_m \in \Re$	第 m 入力変数と出力変数の相関係数
$T = [\boldsymbol{t}_1 \quad \boldsymbol{t}_2 \quad \cdots \quad \boldsymbol{t}_K] \in \Re^{N \times K}$	PLS / 局所 PLS の潜在変数行列
$\boldsymbol{t}_k = [t_{1k} \quad t_{2k} \quad \cdots \quad t_{Nk}]^{\mathrm{T}} \in \Re^N$	潜在変数行列 $\boldsymbol{T}$ の第 k 列
$\boldsymbol{W} = [\boldsymbol{w}_1 \quad \boldsymbol{w}_2 \quad \cdots \quad \boldsymbol{w}_K] \in \Re^{M \times K}$	PLS / 局所 PLS のローディング行列
$\boldsymbol{w}_k = [w_{1k} \quad w_{2k} \quad \cdots \quad w_{Mk}]^{\mathrm{T}} \in \Re^M$	ローディング行列 $\boldsymbol{W}$ の第 k 列
$\boldsymbol{w} = [w_1 \quad w_2 \quad \cdots \quad w_M]^{\mathrm{T}} \in \Re^M$	SVR におけるウェイトベクトル

$\boldsymbol{X} = [\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_N]^{\mathrm{T}} \in \Re^{N \times M}$	前処理前の入力変数の値を含む行列
$\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{x}}_1 \quad \tilde{\boldsymbol{x}}_2 \quad \cdots \quad \tilde{\boldsymbol{x}}_N]^{\mathrm{T}} \in \Re^{N \times M}$	前処理後の入力変数の値を含む行列
$\boldsymbol{x}_n = [x_{n1} \quad x_{n2} \quad \cdots \quad x_{nM}]^{\mathrm{T}} \in \Re^{M}$	入力変数行列 $\boldsymbol{X}$ の第 n 列（第 n サンプル）
$\tilde{\boldsymbol{x}}_n = [\tilde{x}_{n1} \quad \tilde{x}_{n2} \quad \cdots \quad \tilde{x}_{nM}]^{\mathrm{T}} \in \Re^{M}$	入力変数行列 $\tilde{\boldsymbol{X}}$ の第 n 列（第 n サンプル）
$\boldsymbol{x}_q = [x_{q1} \quad x_{q2} \quad \cdots \quad x_{qM}]^{\mathrm{T}} \in \Re^{M}$	クエリ（出力の推定が要求されたときの入力変数）
$\bar{x}_{*m} \in \Re$	前処理前の第 m 入力変数の平均値
$\hat{x}_{*m} \in \Re$	前処理前の第 m 入力変数の類似度重み付き平均値
$\boldsymbol{y} = [y_1 \quad y_2 \quad \cdots \quad y_N]^{\mathrm{T}} \in \Re^{N}$	前処理前の出力変数の値を含むベクトル
$\tilde{\boldsymbol{y}} = [\tilde{y}_1 \quad \tilde{y}_2 \quad \cdots \quad \tilde{y}_N]^{\mathrm{T}} \in \Re^{N}$	前処理後の出力変数の値を含むベクトル
$\bar{y} \in \Re$	前処理前の出力変数の平均値
$\hat{y} \in \Re$	前処理前の出力変数の類似度重み付き平均値
$\hat{y}_q \in \Re$	クエリ $\boldsymbol{x}_q$ に対する出力の推定値
$\alpha_0 \in \Re$	SVR における定数項
$\boldsymbol{\alpha} = [\alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_N]^{\mathrm{T}} \in \Re^{N}$	SVR における回帰係数ベクトル
$\boldsymbol{\beta} = [\beta_1 \quad \beta_2 \quad \cdots \quad \beta_M]^{\mathrm{T}} \in \Re^{M}$	PLS と LASSO における回帰係数ベクトル
$\gamma \in \Re$	ガウシアンカーネルのパラメータ
$\varepsilon \in \Re$	SVR における許容誤差
$\sigma_{d_n} \in \Re$	d_n（$n = 1, 2, \ldots, N$）の標準偏差

$\sigma_{x_{*m}} \in \Re$	前処理前の第 m 入力変数の標準偏差
$\sigma_y \in \Re$	前処理前の出力変数の標準偏差
$\lambda \in \Re$	LASSO における調整パラメータ
$\varphi \in \Re$	局所 PLS における調整パラメータ
$\boldsymbol{\Omega} = \mathrm{diag}\begin{pmatrix} \omega_1 & \omega_2 & \cdots & \omega_N \end{pmatrix} \in \Re^{N \times N}$	局所 PLS における類似度行列
$\omega_n \in \Re$	第 n サンプルとクエリ $\boldsymbol{x}_q$ の類似度

A.2 手法の説明

A.2.1 前処理

【利用できる方法リスト】

1. 平均化
2. 標準化

平均化は各変数の平均値を 0 にする操作であり，以下のように表現できる．

$$\tilde{x}_{nm} = x_{nm} - \bar{x}_{*m} \tag{1}$$

$$\bar{x}_{*m} = \frac{1}{N}\sum_{n=1}^{N} x_{nm} \tag{2}$$

ここで，$\tilde{x}_{nm}$ は前処理後の変数であり，$\bar{x}_{*m}$ は第 m 入力変数の平均値である．平均化は，推定精度を改善するということよりも，数学的記述やプログラミングを容易にするという目的のために利用される．

標準化は各変数の平均値を 0，分散と標準偏差を 1 にする操作であり，以下のように表現できる．ただし，標準偏差が 0 の変数の値は全て 0 にする．

$$\tilde{x}_{nm} = \frac{x_{nm} - \bar{x}_{*m}}{\sigma_{x_{*m}}} \tag{3}$$

$$\sigma_{x_{*m}} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(x_{nm} - \bar{x}_{*m})^2} \tag{4}$$

ここで，$\sigma_{x_{*m}}$ は第 m 入力変数の標準偏差である．各変数の標準偏差を揃えることで，各変数の単位の違いが出力の推定に影響することを防ぐことができる．

A.2.2　　入力変数選択

【利用できる方法リスト】

1. 全て選択
2. GAPLS
3. VIP
4. LASSO
5. PLS 回帰係数
6. 相関係数 R

GAPLS（Genetic Algorithm-based Partial Least Squares）とは遺伝的アルゴリズム（Genetic Algorithm, GA）を用いた変数選択手法である．GA とは生物の遺伝の様子を模倣した最適化手法である．0 と 1 で表現された染色体に対して突然変異や交叉といった操作を行い新たな染色体を作り出す．そして各染色体について評価値を計算して淘汰・選択を行う．これによって優れた個体の周辺の空間が優先的に探索され，結果として最適に近い解が効率良く発見可能である．GAPLS では染色体の各ビットに $\boldsymbol{X}$ の各変数を割り当て最適な PLS モデルを与える変数の組を探索する．染色体の評価関数としてクロスバリデーションを行った際の決定係数を用いる．これにより予測精度の高いモデルを構築することのできる変数の組み合わせが得られる．

VIP は Variable Influence on Projection もしくは Variable Importance on Projection の略語である．VIP は PLS モデルを用いて各入力変数の重要性（VIP スコア）を評価し，重要性の大きな入力変数を予め設定した数だけ選択する方法であり，第 m 入力変数の VIP スコア V_m は

$$V_m = \sqrt{\frac{M \sum_{r=1}^{R} q_r^2 \boldsymbol{t}_r^{\mathrm{T}} \boldsymbol{t}_r \frac{w_{mr}}{\|\boldsymbol{w}_r\|}}{\sum_{r=1}^{R} q_r^2 \boldsymbol{t}_r^{\mathrm{T}} \boldsymbol{t}_r}} \tag{5}$$

で定義され，潜在変数の数に依存する．q_r，$\boldsymbol{t}_r$，$\boldsymbol{w}_r$ は PLS モデル構築時に導出される．詳しい導出方法は A.2.3 節に掲載されている．

LASSO は Least Absolute Shrinkage and Selection Operator の略語である．LASSO では，最適化問題

$$\min_{\boldsymbol{\beta}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\| + \lambda \|\boldsymbol{\beta}\|_1 \tag{6}$$

の解として得られる回帰係数ベクトル $\boldsymbol{\beta}_{\mathrm{LASSO}}$ の要素が 0 になりやすいという性質を利用する．具体的には，$\boldsymbol{\beta}_{\mathrm{LASSO}}$ の 0 要素に対応する入力変数を除外する．最適化問題中の調整パラメータ λ を調整することで，所望の数の入力変数を選択できる．λ が大きいほど，$\boldsymbol{\beta}_{\mathrm{LASSO}}$ 中の 0 要素の数は増える．ただし，式 $\min_{\boldsymbol{\beta}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\| + \lambda \|\boldsymbol{\beta}\|_1$

(6)で $\boldsymbol{X}$，$\boldsymbol{y}$ の各列の平均は 0 であると仮定している．

PLS 回帰係数を選択した場合，次式で定義される PLS 回帰係数 $\boldsymbol{\beta}_{\mathrm{PLS}}$ の絶対値が大きな要素に対応する入力変数が選択される．選択する入力変数の数は予め設定しておく．

$$\boldsymbol{\beta}_{\mathrm{PLS}} = [\beta_{\mathrm{PLS},1} \quad \beta_{\mathrm{PLS},2} \quad \cdots \quad \beta_{\mathrm{PLS,M}}]^{\mathrm{T}} = \boldsymbol{W}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{q} \tag{7}$$

$\boldsymbol{W}$，$\boldsymbol{P}$，$\boldsymbol{q}$ は PLS モデル構築時に導出される．詳しい導出方法は A.2.3 節に掲載されている．VIP スコアと同様，PLS 回帰係数も潜在変数の数に依存する．回帰係数は各入力変数の出力への影響度合いを表現しているが，ある入力変数を他の入力変数と独立に操作できない場合には，他の入力変数の出力への影響も考慮しなければ，正確な影響度合いを計算することができないことに注意が必要である．

相関係数 R を選択した場合，次式で定義される，第 m 入力変数と出力変数の相関係数 R_m が大きい入力変数が選択される．選択する入力変数の数は予め設定しておく．

$$R_m = \frac{1}{\sigma_{x_{*m}}\sigma_{y_*}} \sum_{n=1}^{N} (x_{nm} - \bar{x}_{*m})(y_n - \bar{y}) \tag{8}$$

$$\bar{y} = \frac{1}{N} \sum_{n=1}^{N} y_n \tag{9}$$

$$\sigma_y = \sqrt{\frac{1}{N-1} \sum_{n=1}^{N} (y_n - \bar{y})^2} \tag{10}$$

A.2.3　モデリング

【利用できる方法リスト】

1. SVR
2. PLS
3. 局所 PLS

SVR（Support Vector Regression）は SVM（Support Vector Machine）を回帰分析へと応用した手法である. SVR においてモデル構築用サンプル中の第 n サンプル $\boldsymbol{x}_n$ に対する出力の推定値は回帰式 f を用いて以下のように表される．

$$f(\boldsymbol{x}_n) = \phi(\boldsymbol{x}_n)^{\mathrm{T}} \boldsymbol{w} + \alpha_0 \tag{11}$$

ここで，ϕ はある非線形関数，$\boldsymbol{w}$ はウェイトベクトル，α_0 は定数項である．SVR においては下記の式を最小化するように学習が行われる．

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N} E_{\varepsilon}\left(f(\boldsymbol{x}_n) - y_n\right) \tag{12}$$

ただし，

$$E_\varepsilon\left(f\left(\boldsymbol{x}_n\right)-y_n\right)=\max\left(0,\left|f\left(\boldsymbol{x}_n\right)-y_n\right|-\varepsilon\right) \tag{13}$$

である．式(12)を最小化することによってモデル構築用サンプルへの当てはまり（式(12)の右の項の最小化）と汎化能力（式(12)の左の項の最小化）とのバランスの取れた非線形回帰モデルが得られる．式(12)の C は誤差項の重みであり 2 つの項の間で重み付けを調整する．式(12)の ε は許容誤差である．出力の推定誤差の絶対値が ε 以内の領域を ε チューブと呼ぶ．$\boldsymbol{w}$ は以下のように与えられる．

$$\boldsymbol{w}=\sum_{n=1}^{N}\alpha_n\boldsymbol{\phi}\left(\boldsymbol{x}_n\right) \tag{14}$$

ただし α_n は式(12)の最小化により決定される．よって式(11)よりクエリ $\boldsymbol{x}_q$ に対する出力の推定値は以下の式で与えられる．

$$\hat{y}_q=f\left(\boldsymbol{x}_q\right)=\sum_{n=1}^{N}\alpha_n K\left(\boldsymbol{x}_q,\boldsymbol{x}_n\right)+\alpha_0 \tag{15}$$

K は

$$K\left(\boldsymbol{x}_q,\boldsymbol{x}_n\right)=\boldsymbol{\phi}\left(\boldsymbol{x}_q\right)^{\mathrm{T}}\boldsymbol{\phi}\left(\boldsymbol{x}_n\right) \tag{16}$$

でありカーネル関数と呼ばれる．本ソフトウェアではカーネル関数として以下のガウシアンカーネルを使用している．

$$K\left(\boldsymbol{x}_q,\boldsymbol{x}_n\right)=\exp\left(-\gamma\left\|\boldsymbol{x}_q-\boldsymbol{x}_n\right\|^2\right) \tag{17}$$

γ はモデル構築用サンプルのカーネルの分散が最大となるように自動的に決定される．

PLS と局所 PLS（Locally Weighted PLS）について説明する．PLS と局所 PLS の大きな違いは，非線形性に対応できるかどうか，および，出力の推定が要求される前に回帰係数を計算しているかどうかにある．PLS モデルは線形モデルであるため，非線形性には対応できない．また，PLS モデルの回帰係数は出力の推定が要求される前に計算できる．一方，局所 PLS ではサンプル間類似度に基づいてサンプルに重みをつけ，出力の推定が要求される条件（クエリ）の付近で局所的な線形モデルを構築するため，非線形性に対応し得る．また，クエリが与えられてから回帰係数を計算する．

局所 PLS モデルによる出力推定のアルゴリズムは下記の通りである．局所 PLS モデルは PLS モデルを含んでおり，アルゴリズム中のサンプル間類似度行列 $\boldsymbol{\Omega}$ を単位行列とすれば，PLS モデルによる出力の推定ができる．

【PLS / 局所 PLS モデルによる出力変数の予測アルゴリズム】

1. 潜在変数の数 R を決定する．
2. $r = 1$ とする．
3. 類似度行列 $\boldsymbol{\Omega} = \mathrm{diag}\left(\omega_1 \quad \omega_2 \quad \cdots \quad \omega_N\right)$ を以下の式で計算する．

$$\omega_n = \exp\left(-\frac{d_n}{\sigma_d}\varphi\right) \tag{18}$$

$$d_n = \sqrt{\left(\boldsymbol{x}_n - \boldsymbol{x}_q\right)^{\mathrm{T}}\left(\boldsymbol{x}_n - \boldsymbol{x}_q\right)} \tag{19}$$

$$\sigma_d = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(d_n - \bar{d})^2} \tag{20}$$

$$\bar{d} = \frac{1}{N}\sum_{n=1}^{N} d_n \tag{21}$$

ここで，d_n はクエリと第 n サンプル間の距離，σ_d は距離の標準偏差，$\bar{d}$ は距離の平均値である．

4. 以下の式で，$\boldsymbol{X}_r$，$\boldsymbol{y}_r$，$\boldsymbol{x}_{qr}$ を計算する．この操作は各変数の類似度重み付きを 0 にすることを意味する．この操作を行うことで，精度の改善が見込める．

$$\boldsymbol{X}_r = \boldsymbol{X} - \boldsymbol{1}_N\left[\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_M\right] \tag{22}$$

$$\boldsymbol{y}_r = \boldsymbol{y} - \boldsymbol{1}_N\hat{y} \tag{23}$$

$$\boldsymbol{x}_{qr} = \boldsymbol{x}_q - \left[\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_M\right]^{\mathrm{T}} \tag{24}$$

$$\hat{x}_m = \sum_{n=1}^{N}\omega_n x_{nm} \Big/ \sum_{n=1}^{N}\omega_n \tag{25}$$

$$\hat{y} = \sum_{n=1}^{N}\omega_n y_n \Big/ \sum_{n=1}^{N}\omega_n \tag{26}$$

ここで，$\boldsymbol{1}_N$ は N 個の 1 を要素とする列ベクトルである．また，$\hat{x}_m$ と $\hat{y}$ は第 m 入力変数および出力変数の類似度重み付き平均値である．

5. 第 r 潜在変数 $\boldsymbol{t}_r$ を計算する．

$$\boldsymbol{t}_r = \boldsymbol{X}_r\boldsymbol{w}_r \tag{27}$$

ここで，$\boldsymbol{w}_r$ は $\boldsymbol{X}_r^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{y}_r\boldsymbol{y}_r^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{X}_r$ の最大固有値に対応する固有値ベクトルである．

6. 第 r 潜在変数に対する再構築ベクトル $\boldsymbol{p}_r$ と回帰係数 q_r を計算する．

$$\boldsymbol{p}_r = \frac{\boldsymbol{X}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{t}_r}{\boldsymbol{t}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{t}_r} \tag{28}$$

$$q_r = \frac{\boldsymbol{y}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{t}_r}{\boldsymbol{t}_r^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{t}_r}. \tag{29}$$

7. クエリ $\boldsymbol{x}_q$ の第 r 潜在変数 t_{qr} を求める．

$$t_{qr} = \boldsymbol{x}_{qr}^{\mathrm{T}} \boldsymbol{w}_r \tag{30}$$

8. $r = R$ であればクエリに対する出力の推定値 $\hat{y}_q$ を計算する．

$$\hat{y}_q = \bar{y} + \sum_{r=1}^{R} t_{qr} q_r \tag{31}$$

そうでなければ，$\boldsymbol{X}_r$，$\boldsymbol{y}_r$，$\boldsymbol{x}_{qr}$ を計算する．

$$\boldsymbol{X}_{r+1} = \boldsymbol{X}_r - \boldsymbol{t}_r \boldsymbol{p}_r^{\mathrm{T}} \tag{32}$$

$$\boldsymbol{y}_{r+1} = \boldsymbol{y}_r - \boldsymbol{t}_r \boldsymbol{q}_r^{\mathrm{T}} \tag{33}$$

$$\boldsymbol{x}_{qr+1} = \boldsymbol{x}_{qr} - t_{qr} \boldsymbol{p}_r^{\mathrm{T}} \tag{34}$$

9. $r = r+1$ と更新し，ステップ 5 に戻る．

付録B　ソフトセンサ構築に関する用語と注意点

本章では，ソフトセンサ構築に関する用語や注意点を説明する．

【オーバーフィッティング（過適合，過学習）】

入力変数の数が過剰であることや，モデルパラメータの設定が不適切であること，サンプル数が少ないことなどにより，モデル構築用サンプルのみに対する推定精度が高いモデルができること．

【アンダーフィッティング】

入力変数の数が過小であることや，モデルパラメータの設定が不適切なことにより，モデル構築用サンプルに対する推定精度が低いモデルができること．

【クロスバリデーション（交差検定）】

クロスバリデーションとは，以下の手順で変数選択手法やモデリング手法のパラメータを決定する方法である．クロスバリデーションを利用すれば，パラメータ調整用データが無くても，パラメータを決定できる．

1. モデル構築用サンプルを選択する．

2. サンプル番号に基づいて，モデル構築用サンプルをいくつかのサブデータに分割する．この際，サブデータに含まれるサンプル数の差が最大でも 1 となるように均等に分割する．また，分割によって得られたサブデータの数をクロスバリデーション分割数と呼ぶ.

3. 入力変数選択とモデリングのパラメータセットの候補を指定する.

4. 3 で設定したパラメータセットの 1 つを選択する．ただし，これまで選択されたパラメータセットとは異なるものを選択する．これまでの手続きで，全てのパラメータセットが選択されている場合，ステップ 11 に進む.

5. サブデータの 1 つをモデル構築用サンプルから除外する.

6. 除外したサブデータ以外のデータとステップ 4 で選択したパラメータの値を利用して，入力変数の選択とモデリングを行う.

7. ステップ 5 で構築したソフトセンサを利用して，ステップ 5 で除外したサブデータに対する出力の推定値と推定誤差を計算する.

8. 全てのサブデータが 1 回ずつ除外されるように，除外するサブデータを変更しながら，ステップ 6 と 7 を繰り返す.

9. ステップ 6 と 7 で得られた推定誤差の合計（クロスバリデーション誤差）を求める.

10. ステップ 4 に戻る.

11. ステップ 9 で計算されるクロスバリデーション誤差の最小値に対応するパラメータセットを採用する.

【モデル構築用サンプル】

モデル構築用サンプルとは，変数選択の指標や回帰係数を計算するために利用されるサンプルである．計算を実行するには，モデル構築用サンプルだけでなく，利用する変数選択手法やモデリング手法に応じたパラメータの指定も必要である．モデル構築用サンプルに合わせてモデルを構築することをフィッティングと呼び，モデル構築用サンプルに対する出力の推定誤差をフィッティング誤差と呼ぶ.

【パラメータ調整用サンプル】

パラメータ調整用サンプルとは，変数選択手法やモデリング手法のパラメータを決定するために利用されるサンプルである．パラメータ調整用サンプルを利用したパラメータ決定手順は下記の通りである.

1. 入力変数選択とモデリングのパラメータセットの候補を指定する.
2. ステップ 1 で設定したパラメータセットの 1 つを選択する．ただし，これまで選択されたパラメータセットとは異なるものを選択する．これまでの手続きで，全てのパラメータセットが選択されている場合，ステップ 6 に進む.
3. モデル構築用サンプルとステップ 2 で選択したパラメータの値を利用して，ソフトセンサを構築する.

4. ステップ 3 で構築したソフトセンサを利用して，パラメータ調整用サンプルに対する出力の推定値と推定誤差（パラメータ調整誤差）を計算する．
5. ステップ 2 に戻る．
6. ステップ 4 で計算されるパラメータ調整誤差の最小値に対応するパラメータセットを採用する．

【モデル検証用サンプル】

モデル検証用サンプルとは，回帰係数などの計算や，パラメータ調整時には利用しないサンプルである．モデル構築用サンプルを利用して構築したモデルが，モデル検証用サンプルに対する出力の推定値を正確に推定できれば，モデルの実装時にも高い推定精度が得られることが期待できる．

【モデル構築用サンプル，パラメータ調整用サンプル，モデル検証用サンプルの選択指針】

モデル構築用サンプル，パラメータ調整用サンプル，モデル検証用サンプルの選択を，適切かつシステマティックに行う方法は確立されておらず，利用者の試行錯誤が必要となる場合が多い．ここでは，一般的な指針を述べる．

1. 各サンプル集合には多様なサンプルが含まれている方が良い．
2. 各サンプル集合に含まれるサンプルの数や多様性が十分でないと，ソフトセンサ実装時に誤差が想定以上に大きくなる可能性が高くなる．
3. モデル構築用サンプルはソフトセンサ実装時にも利用されるため，３つのサンプル集合の中で最も重要であり，優先的に内容を充実させるべきである．
4. 特別な事情がなければ，パラメータ調整用サンプルは空集合とし，クロスバリデーションによってパラメータ決定を行うことを推奨する．これは，モデル構築用サンプルに含まれるサンプル数を多くするためである．パラメータ調整用サンプルを使う場合に比べ，クロスバリデーションでパラメータを決定すると，オーバーフィッティングが起こりやすいが，クロスバリデーション分割数を大きくすることで対応できる．
5. モデル構築の初期段階では，狭い範囲のサンプルのみを利用すると結果の解釈がしやすくなる．例えば，プロセスが定常状態にあると考えられる時のサンプルなどを利用することは有効である場合が多いと考えられる．

【ダイナミックモデル】

ダイナミックモデルとは測定時刻の異なる入力変数を利用して，出力変数を利用する

モデルのことである．本ソフトウェアでは，下記のようにしてダイナミックモデル用のデータを作成する．

1. 入力変数行列 $\boldsymbol{X}_{\mathrm{all}}$ と出力変数ベクトル $\boldsymbol{y}_{\mathrm{all}}$ を準備する．行列，ベクトルに欠損値が含まれていても良い．測定間隔は一定でなければならない．欠損値を含めたサンプル数を N_{all} とする．
2. ソフトの GUI から，ダイナミックの設定を読み込む．ここでは，[0 t_1 t_2 … t_J]と入力されているとする．ここで，t_j は N_{all} より小さい自然数である．
3. 入力変数行列 $\boldsymbol{X}_{\mathrm{all}}$ の 1 行目から t_1 行目を削除し，t_1 行の欠損値を削除後の行列に追加する．
4. ステップ 3 を全ての t_j に対して繰り返す．
5. ステップ 3, 4 で生成した行列と元々の行列を横に並べて新たな入力変数行列 $\boldsymbol{X}_{\mathrm{all,dyn}}$ を作成する．
6. $\boldsymbol{X}_{\mathrm{all,dyn}}$ と $\boldsymbol{y}_{\mathrm{all}}$ のいずれかに欠損値が含む行のデータを削除し，いずれにも欠損値が含まれていない行のみを残すことで，ダイナミックモデル用の入力変数行列 $\boldsymbol{X}$ と出力変数ベクトル $\boldsymbol{y}$ を作成する．

表 A.1 で定義したサンプル数 N は上記作業後の出力変数ベクトルの要素数であり，N_{all} とは必ずしも一致しないことに注意が必要である．