大量の欠損を含む製造データから不良要因を同定するスパースモデリング技術

Sparse Modeling Algorithm Allowing Identification of Failure Factors Using Manufacturing Big Data with High Missing Rate

髙田　正彬 TAKADA Masaaki　　西川　武一郎 NISHIKAWA Takeichiro

近年，プラントや工場では，製造ビッグデータの活用により，製品不良の要因を特定して歩留まりの向上を図っている。しかし，サンプリング検査などでデータに多くの欠損値が含まれると，要因解析が計算量・精度の両面で困難になる場合がある。

東芝グループは，大学共同利用機関法人 情報・システム研究機構 統計数理研究所（以下，統計数理研究所と略記）と共同で，自動的に変数を選択しながら回帰モデルを推定するスパースモデリング技術をベースに，多くの欠損値が含まれるデータから高速・高精度に要因解析を行えるHMLassoという手法を開発した。その結果，これまでの先端的な手法と比べて推定誤差を約41％削減できることを確認した。この手法により，多くの欠損を含んでいても不良要因の同定が可能となり，製造現場の生産性・歩留まり・信頼性の向上が期待できる。

Manufacturing industries have recently been focusing on improving production yield by identifying the causes of product defects through effective utilization of the large volumes of data accumulated in plants and factories, referred to as manufacturing big data. However, it has become difficult to implement factor analysis due to the increasing incidence of data containing many missing values as a result of sampling inspections and other reasons, resulting in both increased computational costs and decreased statistical accuracy.

The Toshiba Group, in cooperation with the Institute of Statistical Mathematics, has developed a new sparse modeling algorithm called HMLasso capable of performing accurate and high-speed factor analysis using manufacturing big data with a high missing rate. Numerical experiments using synthetic data have verified that this algorithm reduces the estimation error by about 41% compared with that of other state-of-the-art methods. The introduction of HMLasso is expected to improve the productivity, yield, and reliability of manufacturing sites.

1. まえがき

製造業では，工場の歩留まりや製品の品質は，売り上げやコストに直結する重要な経営指標である。そのため，これらの指標を向上させることは非常に重要であり，古くから様々な品質管理技法が開発されてきた。近年，IoT（Internet of Things）によるデータ収集範囲の拡大や，製造プロセスデータベースの整備，統計・機械学習技術の高度化などを受けて，自動的・網羅的に品質を解析することが可能となってきた。特に，製造ビッグデータは，データ項目の多い高次元データになることが多く，スパースモデリングと呼ばれる高次元データ解析手法が有効である。

ここでは，スパースモデリング技術の概要と，製造ビッグデータにおける適用課題を解決するために東芝グループが開発したHMLassoという手法について述べる。

2. 製造ビッグデータとスパースモデリング

2.1 製造データにおける統計解析

回帰モデリングは，製造データにおける典型的な統計解析手法である。回帰モデリングとは，対象とする項目（目的変数）を，それ以外の項目（説明変数）で表す式（回帰モデル）を推定することである。具体的には，製造物$i=1, \cdots, n$，データ項目$j=1, \cdots, p$に対して，目的変数をY_i，説明変数を$X_{i1}, \cdots, X_{ip}$として，以下の式(1)で表される。

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \quad (1)$$

回帰モデルは，回帰係数$\beta_0, \beta_1, \cdots, \beta_p$によって容易に解釈できることから，現象の理解を志向したモデリングに特に有用である。例えば，製造データの解析では，目的変数に重要な品質特性を指定し，説明変数に各種センサー値や加工条件を指定することで，品質特性のばらつきを説明する回帰モデルを推定する。その回帰係数は，各種センサーや加工条件が品質特性に与える影響度を表しているため，回帰モデルを用いて品質特性のばらつき要因を同定したり，品質ばらつき低減のための制御に利用したりできる。製造データの回帰モデリングに基づく不良要因解析のイメージを，**図1**に示す。

近年の製造データの特徴は，データ項目が非常に多い高

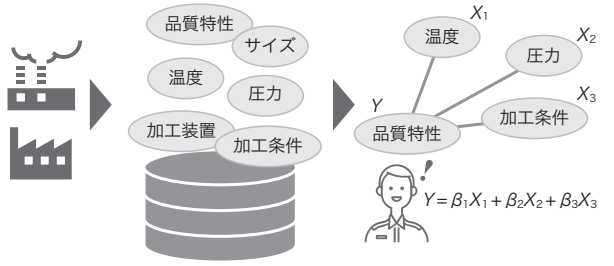


図1. 製造データにおける回帰モデリングに基づく不良要因解析

工場やプラントなどに蓄積される製造履歴データを用いて，品質特性のばらつきを回帰モデリングにより解析する。

Analysis of failure factors based on regression modeling using manufacturing data

次元データであることである。データ収集範囲の拡大やデータベースの整備が進むにつれて，解析に利用できるデータ項目が非常に多くなっている。データ項目が多くなると使える情報量が増えるため，望ましいことであるが，ノイズとなる情報も増え，解析が困難になる。ここで問題となるのは，①データ項目が多過ぎると技術者が全て把握しきれないこと，②手元のデータにはよく当てはまるが将来のデータには当てはまらないモデルが推定されてしまうことであり，①は解釈性，②は精度の問題である。そこで，データ項目を絞り込むことが必要となる。

データ項目を絞り込む方法は，大きく分けて二つある。一つは，ドメイン知識に基づいて関連するプロセスやデータ項目を絞り込む方法である。これは，プロセスの知見や物理法則から，明らかに影響しないデータ項目が除外できる場合に有効である。もう一つは，データ傾向に基づいて変数選択を行う方法である。この方法は，データドリブンで知見を発見したい場合に有効であり，ここでの主題となる。これら二つは排他的ではなく，一般には両者を用いてデータ項目を絞り込む。

2.2 スパースモデリング

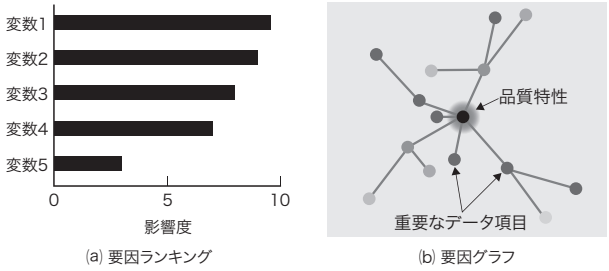
スパースモデリングは，高次元データから自動的に変数を選択しながら回帰モデルを推定する方法である。スパースモデリングでは，多くの項目の中で，実際に品質に関連のある項目は少ない（スパースである）と仮定して，少ないデータ項目で回帰モデルを推定する。この推定の定式化を工夫することで，“変数選択とモデル化を同時に実行する”ことが，スパースモデリングのポイントである。

スパースモデリングは，1996年に提案されたLasso（Least Absolute Shrinkage and Selection Operator）[(1)]にはじまり，今日に至るまで多くの研究で理論的・数値的な有効性が示されており[(2)]，インダストリー分野においても高次元データ解析のコア技術であると考えられる。

表1. データに基づく変数選択の手法一覧

Types of feature selection methods based on data

項　目	説　明	回帰モデリングにおける代表的手法
ラッパー型	変数選択とモデル化を順に繰り返す方法	ステップワイズ法
フィルター型	モデル化の前に軽量な変数選択を行う方法	相関によるスクリーニング[(3)]
埋め込み型	モデル化の中に変数選択が組み込まれている方法	スパースモデリング（Lasso[(1)]）



(a) 要因ランキング　(b) 要因グラフ

図2. 不良要因の可視化の例

不良要因候補のランキングを監視したり，グラフ構造でデータ項目の関係性を理解したりすることで，不良要因の同定を支援できる。

Examples of visualized failure factors

スパースモデリングを含め，データに基づく変数選択手法の分類を**表1**に示す。スパースモデリングは，モデル化の中に変数選択が組み込まれている“埋め込み型”に相当し，ほかのラッパー型やフィルター型と比べて精度と安定性に優れている。

2.3 スパースモデリングを用いた製造データ解析

製造データ解析でスパースモデリングを活用すれば，要因同定作業が網羅的・自動的（又は半自動的）に実施できるようになる。例えば，品質特性のばらつき要因解析では，数百から数万程度の要因候補の中から，数十程度の要因を自動的に絞り込める。また，絞り込まれた要因に対し，回帰係数から影響度を算出すれば，影響度のランキングを作成できる。更に，品質特性に影響のあるデータ項目やデータ項目間の依存関係をグラフ構造で可視化すれば，要因の直感的な理解を促し，要因見逃しの抑制を図れる。可視化のイメージを，**図2**に示す。

3. 新しいスパースモデリング手法

3.1 欠損値問題

製造データは，高次元データであることのほかに，欠損が多く含まれることが特徴である。これは，主にサンプリング

（抜き取り）検査によるものである。例えば，10個体から1個体を抜き出してセンサーで計測する系では，残りの9個体が欠損となるため，欠損率は90％になる。また，抜き取りを行うサンプルは，工程ごとに独立に選択する場合が多い。そのため，工程をまたいで（あるいは全工程で）解析する場合に，データ項目の全ての値がそろっているサンプルは少ない。したがって，欠損を含むサンプルを除外すると，使えるものがほとんどなくなってしまう。

このような場合に欠損を扱う方法は，大きく二つある（**図3**）。一つは，欠損を補完する方法である。単純には平均値や中央値で補完する方法が考えられるが，推定にバイアスが生じてしまう。また，多重代入法やランダムフォレストを用いた補完方法などもあるが，高次元データでは非常に計算負荷が大きく，適用は困難である。

もう一つの欠損を扱う方法は，データを補完することなく，欠損データから直接回帰モデルを推定する方法である。このような方法として，CoCoLasso（Convex Conditioned Lasso）[4]が提案されている。回帰モデル推定の定式化を，共分散行列を用いた形に変形することで，補完せずに直接推定する方法を実現している。

ところが，製造データのような欠損が非常に多いデータにCoCoLassoを適用すると，うまく回帰モデルが推定できないことが分かった。この方法では，多くの欠損を含むデータ項目が含まれることが想定されておらず，全てのデータ項目を同等に扱っていたためである。

3.2　開発手法：HMLasso

東芝グループは，統計数理研究所と共同で，HMLassoという手法を開発した[5]。HMLassoでは，高欠損である変数が含まれている場合でも，高精度に回帰モデルを推定できる。具体的には，以下の式(2)及び式(3)を用いて回帰モデルを推定する。

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \Sigma_{j,k} \tilde{\Sigma}_{jk} \beta_j \beta_k - \Sigma_j \rho_j^{\mathrm{pair}} \beta_j + \lambda \Sigma_j |\beta_j| \right\} \qquad (2)$$

$$\tilde{\Sigma}_{jk} = \underset{\Sigma \succcurlyeq 0}{\operatorname{argmin}} \left\{ \Sigma_{j,k} \left(R_{jk} \left(\Sigma_{jk} - S_{jk}^{\mathrm{pair}} \right) \right)^2 \right\} \qquad (3)$$

ここで，X_jをXのj番目のデータ項目として，S_{jk}^{pair}とρ_j^{pair}は，それぞれ（X_j, X_k）と（X_j, Y）の変数ペアごとに欠損を除外して計算した共分散であり，実測率R_{jk}は（X_j, X_k）がペアで実測されている（欠損していない）割合を表す。R_{jk}を用いていることが既存手法との大きな違いであり，欠損の大小を考慮した定式化で高精度化を図っている。HMLassoの簡易実装は，OSS（オープンソースソフトウェア）として公開している[6]。

この手法の特長は，以下のとおりである。

(1)　高精度に回帰モデルを推定可能　　CoCoLassoは，欠損率を考慮しないため，欠損率が高い項目に引きずられて全体の推定精度が低下してしまう。一方，HMLassoは，欠損率によって共分散の精度が異なることを考慮した定式化となっているため，欠損率が高い項目があっても全体の推定精度が低下せず，高精度な回帰モデルの推定が可能である。

(2)　欠損値の補完プロセスを省略可能　　CoCoLassoと同様に，欠損値を含むデータから直接回帰モデルを構築可能で，全体の計算時間を短縮できる。

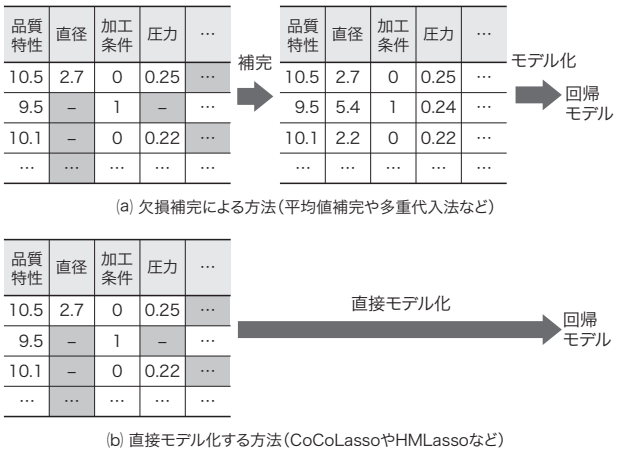
(3)　重要なデータ項目の自動的絞り込み　　Lassoをベースとしているため，品質や歩留まりへの影響度の大きい重要な項目だけを多くのデータ項目から絞り込める。

3.3　手法の有効性

理論・実験の両面で，開発した手法の有効性を確認した。

理論解析では，推定誤差を評価する非漸近理論を用いて，欠損率を活用した開発手法による定式化が，誤差の上限を抑える上で最適となることを示した。これにより，欠損率を考慮しないCoCoLassoよりも理論的に優れていることが，確認できた。

数値実験では，様々な条件の人工データで評価し，平均値補完やCoCoLassoに比べ，ほぼ全ての条件で推定誤差が小さくなることを示した（**図4**）。平均欠損率50％（データ項目によっては，欠損率が90％以上）となるようなデータでは，CoCoLassoに比べ，推定誤差を約41％削減できる



(a) 欠損補完による方法（平均値補完や多重代入法など）

(b) 直接モデル化する方法（CoCoLassoやHMLassoなど）

図3.　欠損データから回帰モデルを推定する方法

開発した手法のHMLassoは，欠損値を補完することなく回帰モデルを推定する。

Methods for estimation of regression model from missing data

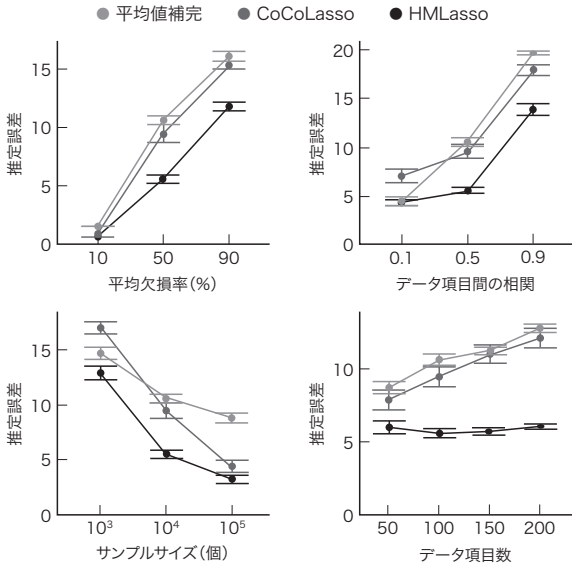


図4. 様々な人工データを用いた各手法の性能比較

サンプルサイズ10^4個，データ項目数100，データ項目間の相関0.5，平均欠損率50％をベースに各条件を変えて実験データを生成し，推定誤差を真の回帰係数と推定した回帰係数の差のL2ノルムで評価した。ほぼ全ての条件で，HMLassoが最も高い精度を達成した。

Results of evaluation of performance of conventional and newly developed methods under various synthetic data conditions

ことを確認した。

4. あとがき

開発した技術を用いることで，大量の欠損を含むデータでも，高い精度で要因解析を行うことが可能となる。

今後，この技術の汎用化・高速化に取り組むとともに，工場・プラントを含む，様々な分野の課題に適用し，生産性・歩留まり・信頼性の向上に貢献していく。

文　献

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

(2) Hastie, T. et al. Statistical learning with sparsity: the lasso and generalizations. CRC press, 2015, 367p.

(3) Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008, **70**, 5, p.849–911.

(4) Datta, A.; Zou, H. CoCoLasso for high-dimensional error-in-variables regression. The Annals of Statistics. 2017, **45**, 6, p.2400–2426.

(5) Takada, M. et al. "HMLasso: lasso with high missing rate". Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19). Macao, China, 2019-08, IJCAI. AAAI Press, 2019, p.3541–3547.

(6) Takada, M. "hmlasso: Lasso with High Missing Rate". CRAN - Package hmlasso. <https://CRAN.R-project.org/package=hmlasso>, (accessed 2020-09-04).

髙田　正彬　TAKADA Masaaki, Ph.D.
研究開発センター 知能化システム研究所
システム AI ラボラトリー
博士（統計科学）　日本統計学会会員
System AI Lab.

西川　武一郎　NISHIKAWA Takeichiro, Ph.D.
研究開発センター 知能化システム研究所
博士（理学）
日本オペレーションズ・リサーチ学会会員
Advanced Intelligent Systems