VARモデルとLassoによる需要予測

Demand Forecasting by VAR Model and Lasso

経営システム工学専攻　坂場賢一郎

1　研究の背景と目的

正確な需要予測は会社経営において重要である．商品の需要予測を行う際，関係のある他の商品の購買記録が参考になる場合がある．しかし，このような商品と商品の関係性が必ずしも明確であるとは限らない．そこで，商品と商品の関連性を発見し考慮することで，需要予測の精度を向上させると推測できる．使用するデータは 165 個の商品の購買記録で，15 年間以上に渡って月ごとに観察された時系列データである．どの商品間に関連性があるのかは不明だが，いくつかの商品の間には関連性が存在することが分かっている．このデータの特性から，複数の系列の間に関係性がある時系列データを扱うことができる VAR モデルを用いる．パラメータは，各式が独立であると仮定することによって，最小二乗法によって推定することが可能だが，今回のデータを用いる場合，最小二乗法のパラメータ推定が不可能である．推定するパラメータの数が各系列の長さを超えるため，最小二乗法の計算に用いる行列が非正則になるためである．この問題に対し解決法を提示することが本研究の目的である．

2　VARモデルとLasso回帰

2.1　K 変量 VAR(p) モデル

VAR モデルは，AR モデルをベクトルに一般化したもので，VAR(p) モデルは $\boldsymbol{y}_t$ を定数と自身の p 期の過去の値に回帰したモデルである．ここで，$y_t = (y_{1t}, \ldots, y_{kt}, \ldots, y_{Kt}), k = 1, \ldots, K$ である．すなわち，

$$\boldsymbol{y}_t = \boldsymbol{c} + \Phi_1 \boldsymbol{y}_{t-1} + \cdots + \Phi_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon_t}$$

というモデルである．ここで，Φ_i は $(K \times K)$ 係数行列で，$i = 1, \ldots, p$, ε_t は K 次元のホワイトノイズである．

K 変量 VAR(p) モデルは K 本の式から成り立ち，各回帰式は，各変数を定数と全変数の p 期間の過去の値に回帰した形となっている．K 変量 VAR(p) モデルが含むパラメータ数を考えると，1 本の回帰式が定数を含め $Kp+1$ 個の係数を持つので，全体では $K(Kp+1)$ 個の係数になる．

2.2　Lasso回帰

Lasso 回帰はパラメータの推定と変数選択を同時に行う手法である．本研究では，3.3 節の問題点を解決することとパラメータの選択による予測精度の向上の目的で使用する．
Lasso 回帰によるパラメータの推定は以下のように表される．

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\| y - \sum_{j=1}^{p} \boldsymbol{\beta_j} \boldsymbol{X_j} \right\|_2^2 \text{ s.t.} \sum_{i=1} |\boldsymbol{\beta_i}| \leq t \quad (1)$$

ここで，t は調整パラメータである．
これをラグランジュの未定乗数法で表すと，

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\| y - \sum_{j=1}^{p} \boldsymbol{\beta_j} \boldsymbol{X_j} \right\|_2^2 + \lambda \sum_{i=1} |\boldsymbol{\beta_i}| \quad (2)$$

となる．λ の値は選択される変数の数に影響を与える．

3　研究手法

3.1　使用データ

本研究で使用するデータは，ある企業から提供されたデータをもとに擬似的に作られた 165 系列の時系列データで，系列長は 186 である．そのデータは，56 種類の商品群 A と，商品群 A の消耗品である 109 種類の商品群 B の購買データで，月ごとに 186ヶ月観測されたものである．

3.2　165 変量 VAR(12) モデル

本研究で扱うデータのいくつかの商品 (系列) の間には関係性があることが分かっている．しかし，どの商品間に関係性があるのかは定かではない．このことか

ら，データを分割せず同時に予測を行う必要があると考える．また，このデータが月ごとに集計されたことに注目して，周期を 1 年 (12ヶ月) とする．つまり，165 変量 VAR(12) モデルを本研究で需要予測に使用するモデルとする．

3.3 165 変量 VAR(12) モデルの係数の推定と問題点

VAR モデルは上記で述べたように複数の式によって成り立つがパラメータの推定の際は，独立に最小二乗法で推定することが出来る (沖本 2013)．今回の場合，165 変量 VAR(12) モデルは 165 個の式によって成り立つため，165 の式は独立に各式のパラメータを求めることが出来るということである．しかし，165 変量 VAR(12) モデルの各式のパラメータ数が系列長よりも大きくなるという問題が生じる．2.1 節より K 変量 VAR(p) のパラメータ数は $K(Kp+1)$ 個で，各式ごとのパラメータ数は式の数である K で割った $(Kp+1)$ 個である．この計算によりより，165 変量 VAR(12) モデルの各式のパラメータ数は 1981 個である．また，各系列の系列長は 186 である．最小二乗法によって求めるパラメータのベクトル $\hat{\boldsymbol{\beta}}$ は $\mathbf{y}$ を目的変数，X を説明変数とした場合，$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$ のように求めるられる．しかし，上記の場合では行列 X が非正則であるため，$(X'X)^{-1}$ の値を算出することが出来ない．しかし，正則化手法の 1 つでパラメータの推定と変数の選択を同時に行うことが出来る Lasso 回帰を用いることで，この問題は解決される．

3.4 データに合った λ の値の選択

上述の問題を解決するために Lasso 回帰を用いて VAR モデルのパラメータを求める．Lasso 回帰を用いるには制約パラメータ λ の値を適当なものに決定しなければならない．この λ の値の大小によって選択される変数の数が変わり，予測値にも変化を及ぼす．λ の値による予測値の変化の例を以下の図 1 に示す．緑の線は実測値，青の線は予測値である．各予測は縦の点線より左にある実測値を用いて予測を行っている．以降，制約パラメータ λ を λ の値と記述する．

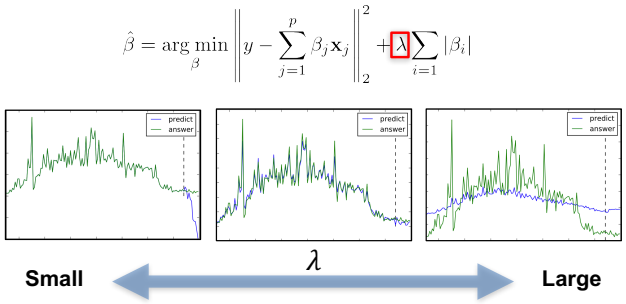


図 1: λ の値による予測の変化

図 1 から，Lasso 回帰で予測をより正確に行うためには適した λ の値を選ぶ必要があることが分かる．しかし，VAR モデルのパラメータを Lasso 回帰で推定する場合，λ の値を推定する適切な方法は，過去の研究等では見受けられない．

そこで，本研究では以下の 6 つの手法を提案する．これらを比較し最も適当な方法を採用する．

1. データから最後の 12ヶ月分を取り除き，数種類の λ で 12ヶ月予測　→　取り除いた期間の後半の 6ヶ月のみの RMSE が小さい λ を選択
2. データから最後の 12ヶ月分を取り除き，数種類の λ で 12ヶ月予測　→　取り除いた期間の後半の 3ヶ月のみの RMSE が小さい λ を選択
3. データから最後の 12ヶ月分を取り除き，数種類の λ で 12ヶ月予測　→　RMSE の小さい λ を選択
4. データから最後の 6ヶ月分を取り除き，数種類の λ で 12ヶ月予測　→　RMSE の小さい λ を選択
5. 数種類の λ で予測　→データのある部分と予測部分の標準偏差の差，実測データとの RMSE の和が小さい λ を選択
6. 数種類の λ で予測　→　AIC を参考とした指標が小さいものを選択

6 つの提案手法の比較方法は，各系列の最後 12ヶ月を取り除き，残りのデータと各提案手法により決定した λ の値を用いて予測を行うことで精度を確かめるという方法を用いる．6 つの提案手法は全て事前に用意している λ の値を逐一用いて指標を算出し，その指標を比較するという方法である．よって，事前に用意している λ の値を全て用いて，取り除いた 12ヶ月を予測し，RMSE を算出することで最小の RMSE を求めることができる．

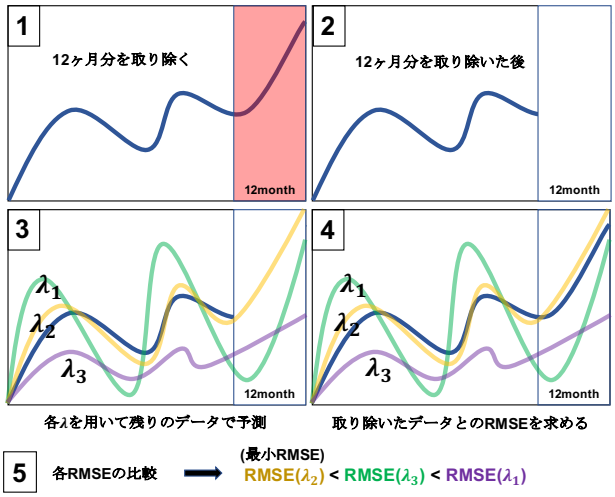


図 2: 最小 RMSE の算出方法

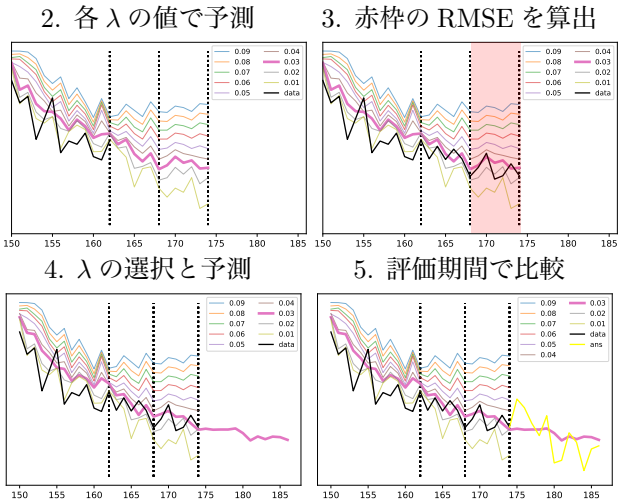


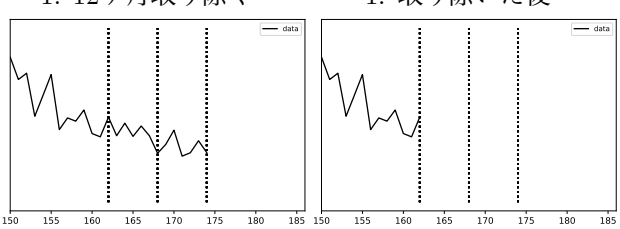
図 3: 提案手法 1 の詳細

本研究では，各手法で決定した λ の値を用いて予測したときの RMSE と事前に求めた最小の RMSE の誤差の比率を計算することで比較を行った．比較結果は以下の表 1 の示す．1 行目「平均の値」は 165 系列の誤差率の平均である．2 行目の「5% 以内」は誤差率が 5% 以内である系列の個数で,「10% 以内」,「15% 以内」も同様である．最後に「0%」は最小 RMSE との誤差が 0% で，すなわち最も適切な λ を選んだ系列の個数である．

表 1: λ の値選択方法の比較結果

	手法 1	手法 2	手法 3	手法 4	手法 5	手法 6
平均の値	25%	27%	29%	32%	63%	200%
5% 以内	69	69	68	61	59	56
10% 以内	89	91	84	74	68	60
15% 以内	103	102	95	84	76	67
0%	56	51	53	1	42	1

提案手法 1 は 1 つの項目を除いて他の手法より結果が優れている．これより，提案手法 1 を採用する．採用された手法の詳細を以下に示す．



3.5 既存手法との比較

Lasso により推定された VAR モデル (LassoVAR) と他の既存手法による予測精度の比較を行った．既存手法には ARIMA モデルと AR モデルを用いた．２つのモデルには R 言語のパッケージである forecast[2] の auto.arima 関数，デフォルトで R 言語の環境に実装されている ar 関数を用いる．また，ラグなどのパラメータは各パッケージ内で自動で最適なものが選択されている．比較手法は，データの最新の 12ヶ月を取り除き，残ったデータを用いて予測を行い RMSE を算出することによって行う．

表 2: モデルごとの RMSE 最小の系列数

LassoVAR	ARIMA	AR	Even
104	40	13	8

3.6 LassoVAR(12) と VAR(12) の精度比較

LassoVAR(12) と VAR(12) の精度の比較をシミュレーションにより行った．通常の VAR モデルは系列数が多いとパラメータの推定が出来ないため，扱う系列数は，1，2，4，6，8，10，12 系列で，系列長は 40，60，80，100，120，140，160 である．扱う系列は互いに相関のある正規化された 134 系列の中からランダムに選ばれている．この条件のもと，12 期先までの予測を各設定で，あらかじめ用意しておいた予測の答えとの誤差を RMSE によって計算し，系列全体での平均をとる操作を 500 回行う．表 3 は LassoVAR(12) と VAR(12) の

結果の差を示しており，LassoVAR モデルによる予測の RMSE から VAR モデルを用いた予測の RMSE を引いた値である．すなわち，値が正であれば LassoVAR(12) モデルの精度が良いということになる．なお，表 3 の「-」で表されている項目は，パラメータの数が系列長を超えているので最小二乗法ではパラメータの値を推定できず，予測が不可能な部分である．

表 3: LassoVAR(12) モデルと VAR(12) モデルの RMSE の差

	40	60	80	100	120	140	160
1	-0.0026	0.0002	0.0150	-0.0016	0.0243	0.0059	0.0146
2	0.0062	0.0115	0.0263	0.0049	0.0080	0.0182	0.0025
4	-	0.0108	0.0213	0.0136	0.0912	0.0187	0.0212
6	-	-	0.0138	0.0017	0.0306	0.0297	0.0237
8	-	-	-	0.0292	0.0356	0.0356	0.0426
10	-	-	-	-	-	0.0286	0.0409
12	-	-	-	-	-	-	0.0325

表 3 の値は全体的に正であるので LassoVAR(12) の予測精度が優れていると考えられる．さらに，12 変量，系列長 40 のような場合では通常の VAR(12) モデルではパラメータを推定できないが，LassoVAR モデルでは可能であるという点でも優れていると考えられる．

4 まとめ

本研究では，VAR モデルの各式の変数の数が系列長を超えるという条件下において，パラメータ推定に最小二乗法を用いることが出来ないという問題点を，Lasso 回帰を用いることによって解決した．この際，VAR モデルの各式は独立であるという仮定を置いている．また，Lasso 回帰を用いる際に決定が必要な制約パラメータの選択手法を 6 つ提案をし，比較を行った．そして，6 つの方法の中から予測に最も適した手法を決定し，これを最終的な提案手法とした．

上述の方法による VAR モデルのパラメータ推定を 165 個の商品の購買データに適用し，予測を行った．このデータの特性からラグを 12 としたため，165 変量 VAR(12) モデルを用いることとした．このデータの系列長は 186，各式の変数の数は 1981 であるため，最小二乗法によるパラメータ推定は不可能である．その結果を，AR モデルと ARIMA モデルを用いた予測と比較し，この 2 つのモデルよりも精度が高いことを確認した．

さらに，VAR モデルのパラメータを Lasso 回帰により求めた場合と最小二乗法により求めた場合の比較も行った．この比較には，最小二乗法でもパラメータ推定が可能な条件下で行った．この結果においても，Lasso 回帰を用いた場合の精度が良いことが確認された．

以上の 2 つの結果から，最小二乗法ではパラメータ推定が不可能な条件でも，Lasso 回帰での推定が可能で，その場合，他の単一系列のみを用いて予測を行う ARIMA モデル， AR モデルよりも予測精度が高いことが推察できる．また，最小二乗法でパラメータ推定が可能な場合でも，Lasso 回帰によるパラメータ推定で予測を行う方が精度が良いことが考えられる．このことから，VAR モデルのパラメータを Lasso 回帰で推定することと，提案手法による Lasso 回帰の制約パラメータ選択方法の有用性を示すことが出来たと考えられる

参考文献

[1] Hamilton, J. D. (1994). *Time series analysis (Vol. 2)*, Princeton: Princeton university press.

[2] Rob J. H,Yeasmin K.(2008). *Automatic Time Series Forecasting: The forecast Package for R*, Journal of Statistical Software,Volume 27, Issue 3.

[3] Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society.Series B (Methodological), 267-288.

[4] Zellner, A. (1962). *An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias*, Journal of the American statistical Association, 57(298), 348-368.

[5] 沖本 竜義. (2013). 経済・ファイナンスデータの計量時系列分析, 朝倉書店.

[6] 「Numpy」*http://www.numpy.org/*（最終アクセス 2018 年 1 月 22 日）

[7] 「Pandas」*https://pandas.pydata.org/*（最終アクセス 2018 年 1 月 22 日）

[8] 「Scikit-learn」*http://scikit-learn.org/stable/#*（最終アクセス 2018 年 1 月 22 日）

[9] 「Matplotlib」*https://matplotlib.org/*（最終アクセス 2018 年 1 月 22 日）