統計的モデリングによる予測・知識発見

川崎 能典　学際統計数理研究系 教授

1　統計的モデリングとは?

データが発生する仕組みを — それが「真の構造」かどうかはひとまず棚上げして — なるべく模倣(mimic)するモデルを，共変量やデータのダイナミクス(過去データへの依存)を使って確率分布で表現し，モデル構造や確率分布の定式化探索を情報量規準で行うことを，統計的モデリングと呼びます．

2　財務データの統計的モデリング

生産あるいは売上高は，経済学的には1) 労働投入，2) 資本投入，3) その他の要因(技術進歩や景況感など)で説明できるとされています．ここでは東証プライム市場に上場している企業(1,137社)に関し，2021会計年度のデータを用いて，売上高を高精度に説明するモデルを構築します．

被説明変数である売上高，説明変数群も対数変換後に線形回帰モデルをあてはめる前提で，探索的に(徐々に適合度を改善する形で)データ解析を進めると，1) 誤差分布は歪対称t分布(歪んでいて裾が重い分布)を仮定したモデルが最も良い適合度を与えること，2) モデルの定数項として現れる「その他の要因」は業種ごとに個別に与えた方が全体として良いモデルとなること，3) 幾つかの業種要因は適合度の観点から併合可能であることがわかります．

興味深いのは，業種要因として推定された値の高低は，コロナ禍2年目の経済状況と符合する解釈が可能であることです．本研究は，統計数理研究所公募型共同利用(2023-ISMCRP-2017)に基づく成果です．

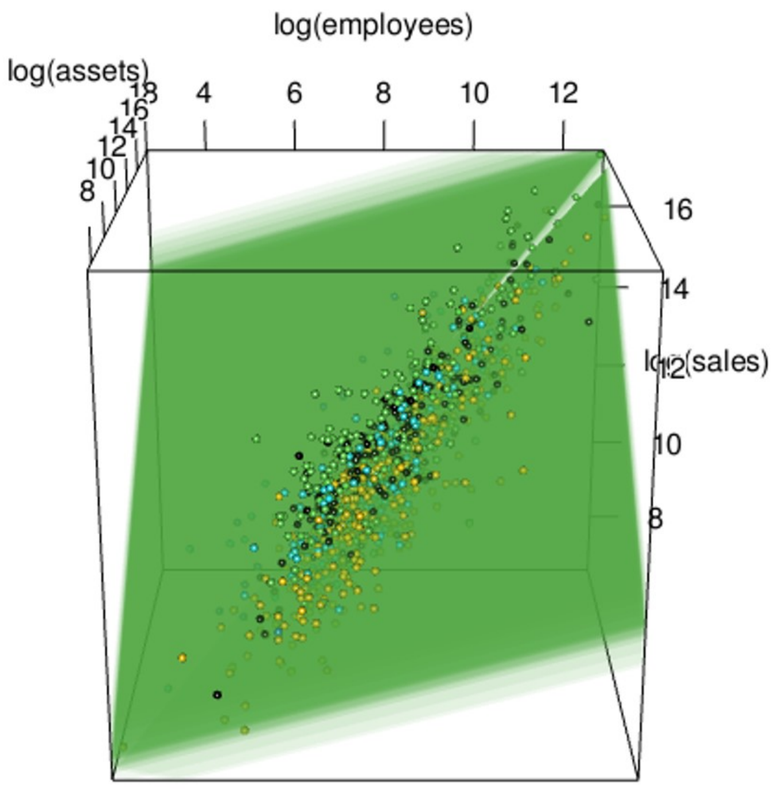

図 1: 業種ダミーと歪対称t分布で推定された最終モデルを可視化

Jimichi. M., Kawasaki, Y., Miyamoto, D., Saka, C. and Nagata, S. (2023), Statistical Modeling of Financial Data with Skew-Symmetric Error Distributions, *Symmetry*, **15**, 1772, doi: 10.3390/sym15091772.

3　金融データのモデリングとリスク管理

金融資産を運用する立場からは，市場価格の変動は収益の源泉である一方で，損失の原因でもあります．実践的には，変動性を収益率(価格の伸縮率)の分散(散らばりの大きさ)と捉え，その時間推移を模倣する時系列モデル(代表的なモデルとしてはGARCHモデル)を過去データから推定し，予測と制御に用います．

しかし，このやり方では，過去データにたまたま大きな変動が含まれていなかった場合は，将来起こりうる大きな下方変動の可能性を過小評価することにつながります．そこで，時系列モデルのあてはめ後に基準化した残差に統計的極値理論を適用することで，データが得られていない分布の裾部分で損失が発生する確率を織り込むことができます．

極値理論の利用に関しては，20年以上特定のモデルが使われ続けてきましたが，我々の研究では，順序統計量を使う古典的な方法が持つバイアスを補正するやり方を採用し，金融リスク管理の文脈に適用しました．図2で赤いカーブ(従来の方法)の上側に青いカーブ(我々の方法)が位置しており，バックテストの観点から青いカーブが高精度であることが(別途検証により)示されます．本研究は，総合研究大学院大学統計科学専攻・貝淵響氏(2022年9月修了)と，Gilles Stupfler教授(仏Angers大学，当時ENSAI)との共同研究による成果です．

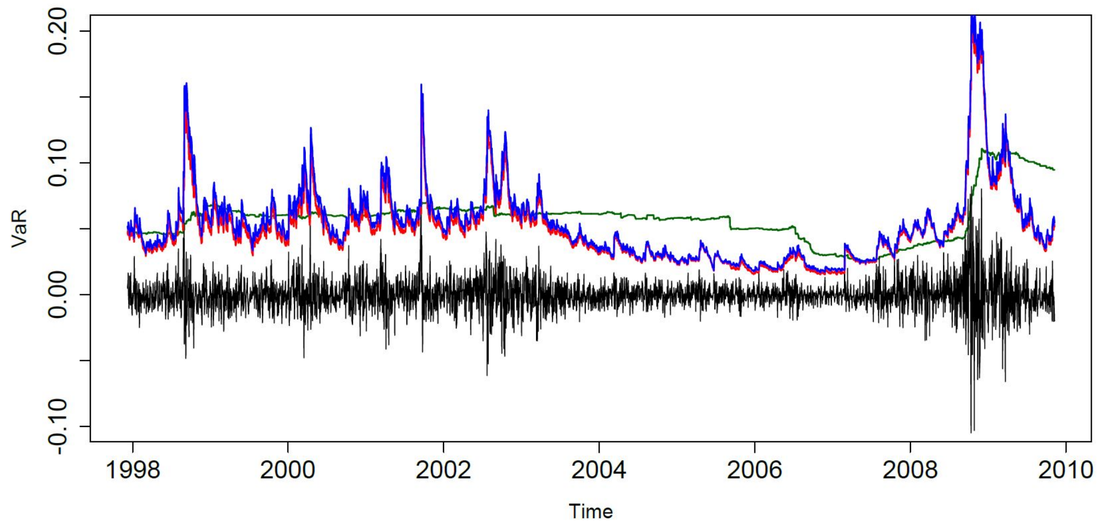

図 2: ダウジョーンズ指数損失率に対する99.9% VaR外挿予測

Kaibuchi, H., Kawasaki, Y. and Stupfler, G. (2022), GARCH-UGH: A bias-reduced approach for dynamic extreme Value-at-Risk estimation in financial time series, *Quantitative Finance*, **22**(7), 1277–1294.

4　テキスト系列からの情報抽出と時系列予測

自然言語処理では，文書(例えばある日の新聞記事)が多数の話題(トピック)から構成されていると考え，トピックはどのような単語がどういう頻度で含まれているかで特徴付けられる，と見なすモデルがあります．この考え方を時間軸方向に拡張したモデルも存在し，さまざまな時間スケール(例えば前日，2日前，3日前...)で単語分布は異なっていて，それらの加重和から各時点のトピックの出方も変わる，というモデリングが可能になります．

これによって，時刻tで第d文書に含まれるトピックiの比率が推定でき，それらを一日内の文書で足し上げることで，時刻tにおけるトピックiのスコアが計算でき，それを様々なtで繰り返すことで，トピックスコア時系列が算出できます．それらを金融資産の変動性(ボラティリティ)を表現したモデルに説明変数として追加することで，モデルの予測精度を向上できる場合があります．本研究は，統計数理研究所公募型共同利用(H25-J-4202, H26-J-4101, H27-2-2012, H28-2-2011)に基づく成果です．

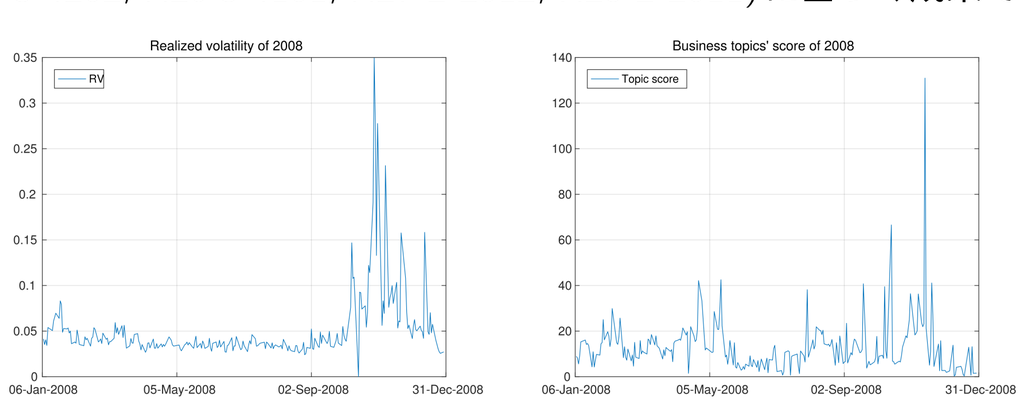

図 3: 推定されたボラティリティ(左)とトピック時系列(右)

Morimoto, T. and Kawasaki, Y. (2017), Forecasting financial market volatility using a dynamic topic model, *Asia-Pacific Financial Markets*, **24**, 149–167.
[日本金融・証券計量・工学学会(JAFEE)2021年度論文賞受賞]