テキスト系列からの情報抽出を利用した時系列予測

川崎　能典　モデリング研究系　教授

概要: テキストデータからトピックを抽出する方法として，潜在ディリクレ分配モデルがよく使われる．テキストデータが毎日得られるような状況では，トピック時系列とでも呼ぶべき系列を推定する方法論に拡張可能である．本研究では，マルチスケール動的トピックモデルの枠組みを用いて新聞記事からトピック時系列を推定し，その中に金融市場の変動性(ボラティリティ)の予測に役立つ系列を抜き出し，予測精度の改善の有無を予測実験で検証する．[本報告は森本孝之氏(関西学院大学理工学部教授)との共同研究である．本文中の文献は Morimoto and Kawasaki (2017) を参照されたい．]

1. マルチスケール動的トピックモデル

単一のテキストが与えられたとき，潜在ディリクレ分配モデル(Latent Dirichlet Allocation Model)はテキストデータからトピックを抽出するよく知られた方法である (Blei et al., 2003)．単語分布は多項分布に従うと仮定し，これが尤度を与える．トピックzは単語分布ϕ_zで特徴付けられ，各文書dは多数のトピックから構成されていると考えるのだが，その組成を表す分布をθ_dとかく．推定はMCMCで行われる．

ここでは，日々更新されるテキストデータを分析しながら，時間軸方向に沿ってトピック時系列を推定したい．そのような方法が，Iwata et al. (2010)によってマルチスケール動的トピックモデルとして提案されている．ここではトピックzの単語分布は時変すなわち$\phi_{t,z}$であるが，それは$t-1$時点でトピックzに関し複数の時間スケールsに基づく単語分布$\hat{\omega}_{t-1,z}^{(s)}$の加重和をパラメータとするディリクレ分布で与える．すなわち

$$\phi_{t,z} \sim \text{Dirichlet}\left(\sum_{s=0}^{S} \lambda_{t,z,s}\hat{\omega}_{t-1,z}^{(s)}\right)$$

である．$\hat{\omega}_{t-1,z}^{(s)}$は例えば$(t-1)-2^{s-1}+1$時点から$t-1$時点までを動く，というような定式化が考えられる．$S=4$だとすれば，$s=3$のときは$t-4$から$t-1$時点までをカバーする単語分布を表す．$s=0$のところでは一様分布と約束しておく．推定のためのMCMCサイクルについてはIwata et al. (2010)を参照．

最終的には，時刻tで第d文書に含まれるトピックiの比率$\theta_{t,d,i}$の推定値から$\text{SC}_t^{(i)} = \sum_{d=1}^{D_t}\theta_{t,d,i}$を構成する．$\text{SC}_t^{(i)}$は時刻$t$におけるトピック$i$のスコア，$D_t$は第$t$日のテキストデータに含まれる文書の数である．

2. データと前処理

テキストデータはロイタージャパンの日本語サイトから記事をスクレイピングした．期間は2008年1月7日から2012年12月28日まで1223日分である．(残念ながらロイターはこのサイトを閉鎖した．)298,205個の文書の中に，所謂ストップワードを除いて24,227語に対象を絞った．トピックスコア系列$\text{SC}_t^{(i)}$は恣意的だが20系列$(i=1,\ldots,20)$抽出した．

テキストに対応させる形で，TOPIXの高頻度データを集約し日次ボラティリティを算出する．第t日の高頻度データから1分刻みの等間隔収益率時系列$r_{t,i}$を生成し，その2乗和(実現ボラティリティ)$\text{RV}_t = \sum_{i=1}^{M} r_{t,i}^2$を日次収益率の代替変数とする．更に，実現quarticity $\text{RQ}_t = (M/3)\sum_{i=1}^{M} r_{t,i}^4$もモデルによっては利用する．

3. HARモデルとその変種による時系列予測

実現ボラティリティの予測の文脈で，その予測性能の高さと推定の容易さから非常に良く使われるようになったのがCorsi (2009)のheterogeneous AR(HAR)モデルである．ここではBollerslev et al. (2016)の定式化に従い，$\text{RV}_{t-j|t-h} = (h+1-j)^{-1}\sum_{i=j}^{h}\text{RV}_{t-i}$(ただし$j \le h$)とし，

$$\text{RV}_t = \beta_0 + \beta_1\text{RV}_{t-1} + \beta_2\text{RV}_{t-1|t-5} + \beta_3\text{RV}_{t-1|t-22} + u_t$$

をHARモデルと定義する．

HARモデルの右辺にトピック時系列SC_tを組み込めば予測が良くなるのではないか，というのがこの研究の仮説である．これをHAR-SCと呼ぶことにする．

$$\text{RV}_t = \beta_0 + \beta_1\text{RV}_{t-1} + \beta_2\text{RV}_{t-1|t-5} + \beta_3\text{RV}_{t-1|t-22} + \gamma\text{SC}_{t-1} + u_t$$

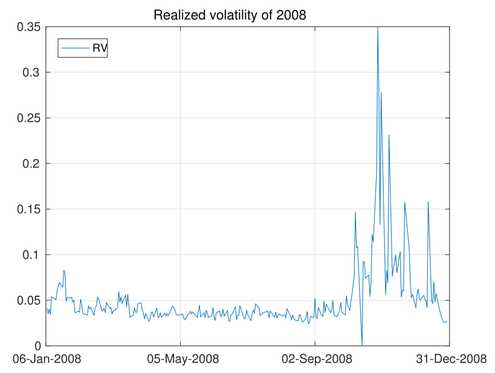
一方，Bollerslev et al. (2016)では，RV_{t-1}の係数をRQ_tに依存させる定式化が提案されていて，予測能力の向上が見られる．これをHARQモデルと呼ぶ．

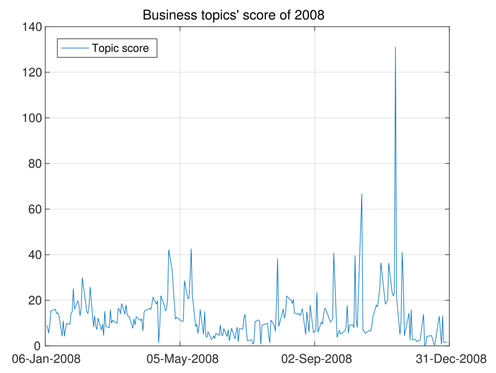
$$\text{RV}_t = \beta_0 + (\beta_1 + \beta_{1Q}\text{RQ}_{t-1}^{1/2})\text{RV}_{t-1} + \beta_2\text{RV}_{t-1|t-5} + \beta_3\text{RV}_{t-1|t-22} + u_t$$

これに対してトピックスコア時系列を右辺に追加するモデルも考えられる．これをHARQ-SCと呼ぶ．

$$\text{RV}_t = \beta_0 + (\beta_1 + \beta_{1Q}\text{RQ}_{t-1}^{1/2})\text{RV}_{t-1} + \beta_2\text{RV}_{t-1|t-5} + \beta_3\text{RV}_{t-1|t-22} + \gamma\text{SC}_{t-1} + u_t$$

Morimoto and Kawasaki (2017)では他のモデルも予測比較の俎上に乗せてはいるが，紙幅の関係でこの4つに絞って結果を報告する．なお，以下の図は左が2008年1月7日から12月29日までのRV_t，右が単語分布からビジネス関連と思われるトピックスコア時系列の推定値である．





4. 予測の実証分析

予測誤差の評価は，ここではMSEとQLIKEを用いる．RV_tに対するモデルの予測値をX_tと書くとき，それぞれ

$$\text{MSE}(\text{RV}_t, X_t) \equiv (\text{RV}_t - X_t)^2$$

$$\text{QLIKE}(\text{RV}_t, X_t) \equiv \frac{\text{RV}_t}{X_t} - \log\left(\frac{\text{RV}_t}{X_t}\right) - 1$$

で定義される(Patton, 2011)．ここでは1期先外挿予測の結果だけを報告する．推定と予測の更新にあたっては，サンプルサイズを400日分に固定して推定ウィンドウをずらしていくやり方(Rolling Window, RW)と，401時点以降使える過去データを全て使ってモデルを再推定するやり方(Increasing Window, IW)の2通りの定式化を試した．

予測結果は以下の通りである．HARモデルをベースラインとしてその誤差関数の値を1に基準化して結果を示している．MSEで評価する場合には，総じてSCを入れた効果よりRQを取り込んだ効果の方が大きいものの，HARQ-SCがRW, IW問わずに良い．QLIKEで評価すると，RQの導入は逆効果となっており，HAR-SCがわずかにHARを凌ぐ．

	HAR	HARQ	HAR-SC	HARQ-SC	SC
MSE(RW)	1.000	0.5562	0.9658	0.5369	$SC^{(11)}$
MSE(IW)	1.000	0.8408	0.9678	0.8175	$SC^{(11)}$
QLIKE(RW)	1.000	1.3781	0.9891	1.3439	$SC^{(3)}$
QLIKE(IW)	1.000	1.1529	0.9883	1.1292	$SC^{(18)}$

謝辞 本研究は統計数理研究所共同利用(H25-J-4202, H26-J-4101, H27-2-2012, H28-2-2011)に基づく成果である．

参考文献

Morimoto, T. and Kawasaki, Y. (2017), Forecasting financial market volatility using a dynamic topic model, *Asia-Pacific Financial Markets*, 24, 149–167.