Prediction of multivariate responses with a select number of principal components

島根大・総合理工　内藤貫太

1　序

マイクロアレイデータに見られるような、高次元少標本における有効な統計解析手法の開発とその評価が重要である。遺伝子発現データの解析では、応答が離散変数である場合が多く、もっぱら判別の問題として定式化され、また実際そのような研究が盛んに行われてきている。一方で、例えば生存時間が応答変数として扱われる場合には、連続変数としての扱いとなり、回帰の枠組みとなる。予測の問題は回帰における基本的課題として大変重要であるが、説明変数が高次元の場合の回帰予測への取組みが始まっている。例えば、Bair et al (2006) では、教師付き主成分分析による 1 変数の予測が提案されている。本研究では、説明変数が高次元の場合における、多変量の予測の方法を提案し、その有効性を調べた。

2　設定

考える多変量回帰予測の設定は、良く知られた多変量回帰モデル

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{1}$$

である。ここで、$\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_N]^T$ は応答変数行列であり、各列 $\mathbf{y}_i$ は $q \times 1$ 列ベクトルである。また、$\mathbf{B} = [\vec{\beta}_1 \cdots \vec{\beta}_p]^T$ は $p \times q$ 回帰係数行列で、$\mathbf{E}$ は $N \times q$ 誤差行列を示す。p 個の説明変数からなる説明変数行列 $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_p]$, の各列 $\mathbf{x}_j$ は中心化されているものとする。ここで、p と N の大小に係わらず用いることのできる予測手法の開発が重要であることに注意しておく。

3　提案方法の概要

提案手法は次のステップを踏む：

3.1　Variable Ranking

p 個の説明変数において、その“ 重要度 ”のランキングを行う。Bair et al (2006) においては、各説明変数と 1 変量応答の相関に基づいてランキングを行っている。本研究では、多変量回帰の枠組みゆえ、説明変数行列と多変量応答の正準相関行列の固有ベクトルに基づいてランキングを行うことを提案している。このようにして得たランキングに基づき、説明変数行列の列をソートし、適当な（最初の）m 列のみを含んだ説明変数行列を作る。

3.2 Dimension Reduction

ソートされた説明変数行列（$N \times m$）に主成分分析（特異値分解）を適用し、適切な H 個の主成分を抽出する。ここで問題は、どのようにその $H = H(m)$ を決めるかであるが、そのためには Koch and Naito (2007) で提案された選択手法を適用する。実際には、m の上限 M を定め、各 $m \leq M$ に対して $H = H(m)$ を Koch and Naito (2007) の基準で選択する。その基準を最大にした (m, H) が以下において用いられる。

3.3 Prediction

ソートした説明変数行列の最初の H 個の主成分で作られる $N \times H$ 行列 $\widetilde{\mathbf{X}}_{m,H}$ により、多変量回帰モデル

$$\mathbf{Y} = \widetilde{\mathbf{X}}_{m,H}\mathbf{B}_r + \mathbf{E} \tag{2}$$

を考え、最小 2 乗法に基づいた従来の回帰予測手法を適用し、予測値を得る。ここで、$\mathbf{B}_r$ は $H \times q$ 回帰係数行列である。

4 提案手法の理解

モデル (1) と (2) の比較からわかるように、“ 説明変数行列をいかにうまく縮小するか？ ”、という点がこの手法のポイントとなる。

5 提案手法の評価

ここで提案された手法は、Bair et al (2006) の多変量への一般化になっていることが示される。マイクロアレイを含む幾つかの実データへの適用を通して、またシミュレーション実験により、提案手法が有効に機能することがわかった。$q = 1$ の場合、つまり従来の重回帰の設定でも、Bair et al (2006) の方法を凌ぐ場合があることを確認できた。

参考文献

[1] Bair, E., Hasie, T., Paul, D. and Tibshirani, R. (2006) Prediction by supervised principal components. *J. Amer. Statist. Assoc.*, **101**, 119-137.

[2] Koch, I. and Naito, K. (2007) Dimension selection for feature selection and dimension reduction with principal and independent component analysis, *Neural Computation*, **19**, 513-545.