高次元データに対する線形判別法の比較

東京理科大・理・院　兵頭　昌

マイクロアレイデータ等に代表されるデータの変数の数がサンプルサイズより大きい高次元データについて最近関心が高まっている．Dudoit et al. (2002, *J. Amer. Statist. Assoc.*, **97**) では 100 よりも少ない腫瘍の標本に対して，5000 ∼ 10000 の遺伝子データが使われている．一般にデータの次元 p がサンプルサイズ n よりも大きくなると，標本共分散行列が退化し，線形判別関数が定義できないという問題が生じる．本報告では，このような状況における線形判別法の補正法を紹介しそれらの誤判別確率の比較を行った．比較に関する報告は，前例がいくつかある．Dudoit et al. (2002) において腫瘍の遺伝子データに対して様々な判別手法の比較を行い，その結果から対角線形判別法が他の複雑な手法より誤判別率が低くなることを報告している．また，対角線形判別法を用いた際の誤判別確率の近似を Bickel and Levina (2004, *Bernoulli*, **10**) が導出している．ムーア・ペンローズ逆行列を用いた判別法は，Xu et al. (2008, *Computational Statistics and Data Analysis.*, doi:10.1016/j.csda.2008.02.005) におけるシミュレーション結果から誤判別確率が高くなることが報告されている．また, 兵頭, 山田 (2008, 2008 年度統計関連学会連合大会講演報告集, 38) ではムーア・ペンローズ逆行列を用いた線形判別法の誤判別確率の近似を与えている．その結果から, 誤判別確率はトータルサンプルサイズに対する各群のサンプルサイズの割合に依存することがわかった．Xu et al. (2008) で提案される判別法では Ledoit and Wolf (2005, *J. Multivariate Anal.*, **88**) によって導出された推定量を用いて線形判別関数を定義している．この判別法は，シミュレーションによる比較からムーアペンローズ逆行列を用いた判別法や対角線形判別法よりも誤判別確率が低くなることが報告されている．また，二次判別においては Kubokawa and Srivastava (2008, *J. Multivariate Anal.*, doi:10.1016/j.jmva.2008.01.016) で Σ の経験ベイズ推定量を導出しそれを用いた判別基準が提案されており誤判別確率の比較が行われている．その結果から，二次判別において Kubokawa and Srivastava (2008) で導出された経験ベイズ推定量を用いた判別基準は他の判別基準に比べて誤判別確率が低くなることが報告されている．そこで本報告では，対角線形判別法，ムーア・ペンローズ逆行列を用いた線形判別関数，Xu et al. (2008) で提案される線形判別法, Srivastava and Kubokawa (2006, *J. Multivariate Anal.*, **73**) で導出された母集団共分散行列 Σ の経験ベイズ推定量を用いた線形判別関数, Kubokawa and Srivastava (2008) で導出された母集団共分散行列 Σ の経験ベイズ推定量を用いた線形判別関数，Bickel and Levina (2008, *The Annals of Statistics*, **36**) で導出されたバンディング推定量を用いた線形判別関数を紹介し比較を行った．

シミュレーションについて

分散共分散行列が等しく平均の異なる 2 つの正規母集団より得られるトレーニングデータを用いて，線形判別関数を定義する．そして，第 2 群から得られるデータを 1000 個抽出しそれらを各線形判別法によって判別しそれぞれの判別法での誤判別確率（本報告では第 2 群から得られるデータを誤って 1 群と判別する確率）の比較を行った．尚，トレーニングデー

タの合計数を 120，次元の数を 100 とした次元の数とトレーニングデータの合計が同程度である設定とトレーニングデータの合計数を 50, 次元の数を 100 とした次元の数がトレーニングデータの合計を上回る 2 つの設定を考えた．また，共分散構造による違いを見るため母分散共分散が AR(1), AR(1)-不均質，一様構造，単位行列の場合を扱った．

シミュレーションより得られた知見

シミュレーションから，共分散行列の構造に大きく影響を受けるものは Bickel and Levina (2008) で導出されたバンディング推定量を用いた線形判別法であった．原因は，この 推定量がある母数の族で一様に良いとされる推定量であることが考えられる．また次元数とトレーニングデータの合計数に関しても各判別法の良し悪しを見ることができた．次元の数とトレーニングデータが同程度の場合対角線形判別法，Xu et al. (2008) で提案される線形判別法，Srivastava and Kubokawa (2006) で導出された推定量を用いた線形判別法が他の判別法より誤判別確率が低い傾向にあった．対して，次元の数がトレーニングデータの数の合計を上回る状況では Kubokawa and Srivastava (2008) で導出された推定量を用いた線形判別法，対角線形判別法，Xu et al. (2008)，Srivastava and Kubokawa (2006) で導出された推定量を用いた線形判別法が他の判別法より誤判別確率が低い傾向にあった．尚，先にも述べたが Bickel and Levina (2008) で導出されたバンディング推定量を用いた線形判別法は共分散行列の構造に大きく影響を受けるが共分散行列が AR(1), 一様構造，単位行列であるとき対角線形判別法と似た挙動を示していた．また，ムーア・ペンローズ逆行列を用いた線形判別法はどの場合も他の判別法に比べ誤判別確率が高くなる傾向にあった．

今後の課題について

本報告では，高次元データに対する補正された線形判別関数を紹介しそれらの誤判別確率を比較した．しかしながら，シミュレーションでは完全にどの判別法が良いかという結論には至らなかった．というのも，本報告でのシミュレーションでは一回の標本抽出によって判別機を得た場合の比較となるため乱数による部分があるためである．今後は，その部分を解消したシミュレーションを行っていきたい．

また，Bickel and Levina (2008) で導出されたバンディング推定量を用いた線形判別法は共分散行列の構造に強く影響を受けるものの AR(1), 一様構造では他の判別法よりも誤判別確率を低くなる場合もあった．今後は，この判別法の誤判別確率の近似を考えそこからこの判別法の特徴を得ることも目標である．