氏名（本籍）	渡邉弘己（群馬県）
学位の種類	博士（理学）
学位記番号	乙第1198号
学位授与の日付	2019年9月30日
学位授与の要件	学位規則第4条第2項該当
学位論文題目	**Estimation of misclassification probabilty for Euclidean distance in high-dimensional data** **（高次元データにおけるユークリッド距離に基づく誤判別確率の推定）**
論文審査委員	（主査）教授　瀬尾　隆 教授　橋口　博樹　　教授　宮岡　悦良 教授　眞田　克典　　准教授　黒沢　健

論文内容の要旨

本論文では、高次元データにおける誤判別確率の推定問題について述べている。

第１章では、古典的な多変量解析法の一つである線形判別分析を導入し、誤判別確率を調べる必要性について、分類精度の面から述べている。また、線形判別分析の理論的な性質を導くための前提条件は「次元数よりも標本サイズの方が大きい。」、「正規性の仮定が必要。」、「分散共分散行列が等しい。」である。しかし、近年では、ゲノムデータのように次元数が標本サイズよりもずっと大きいデータを扱う機会が増えているため、これらの前提条件の下でデータ分析をすることは非現実的である。そこで、高次元データのための判別分析の先行研究について述べている。分散共分散行列が等しい場合の先行研究は、Saranadasa (1993)による一元配置多変量分散分析に基づく方法、Bickel and Levina (2004)による対角線形判別ルール等について述べ、分散共分散行列が異なる場合の先行研究として、Dudoit et al.(2002)による各群の分散共分散行列の対角成分のみを使って逆行列を構成する方法、Chan and Hall(2009)の距離ベースの判別法等について述べている。また、第１章の終わりでは、本論文の第２章から第６章までの大まかな流れ及び概略について述べている。

第２章では、古典的な線形判別分析として、ベイズルールによって構成される判別ルールと、その誤判別確率について紹介している。ベイズルールでは、各群の条件付き密度関数が多変量正規分布に従うという仮定の下、誤判別確率と出現率と判別によるコストから成る期待損失を最小化することを考える。また、最小化した損失には未知パラメータが含まれているため、これらの未知パラメータをトレーニングデータから推定する必要がある。未知パラメータは最尤推定量をプラグイン推定することで得る。そこで得られた判別ルールは線形判別ルールと呼ばれており、フィッシャーの線形判別と本質的に同様の判別ルールが得られる。また、線形判別分析の誤判別確率に関する先行研究についても述べている。Okamoto(1963,1968)では、誤判別確率の大標本漸近展開を $O(n^{-2})$の項まで与え、Siotani and Wang(1977)では、これをさらに $O(n^{-3})$の項まで拡張している。さらに、次元数が標本サイズを超えないが、次元数も標本サイズも大きい場合の漸近展開が Lachenbruch(1968)で与えられており、Fujikoshi(2000)では期待誤判別確率の漸近近似の誤差限界の式を与えている。

第３章では、高次元データにおける判別分析について述べ、これに関する先行研究を紹介している。古典的な多変量解析では、母集団分布が多変量正規分布に従っており、全ての分散共分散行列が等しいという仮定の下で議論されていることが多い。しかし、高次元データが多変量正規分布に従っているという仮定を確かめることはとても難しい問題であり、分散共分散行列が等しいという仮定についても次元数が大きくなればなるほど厳しい仮定である。したがって、高次元データ解析をするにあたり、これらの仮定を置くことは非現実的である。そのため、先行研究では、多変量正規分布に従うという仮定を緩めた仮定を用いて議論している。例えば、Aoshima and Yata(2014)では、Chan and Hall(2009)で得られたスケールアジャストタイプの距離ベースの分類法を基にし、ある仮定の下で誤判別確率が０に収束することを示し、また、別の仮定の下で、誤判別確率の漸近正規性を示している。

第４章では、２群の場合の高次元データにおける誤判別確率の推定問題について述べている。Watanabe et al.(2019)では、Aoshima and Yata(2014)で用いられている、ユークリッド距離を用いた判別関数を用いて、新たな仮定の下で、判別関数の漸近正規性をマルチンゲール中心極限定理を用いて証明している。この漸近正規性を利用して得られる誤判別確率の漸近近似式を得ることに成功している。しかし、この近似式には未知パラメータが含まれているため、それらのパラメータを推定する必要がある。含まれている未知パラメータには先行研究で推定量が与えられていないパラメータが含まれているため、その不偏推定量について、Watanabe et al.(2019)で導出し、その推定量の一致性を証明した。さらに、Watanabe et al.(2019)では、誤判別確率の推定量として、クロスバリデーションを用

いた推定量についても紹介し、その一致性についても証明している。第４章の最後には、第４章に関する数値実験とその結果について述べている。より具体的には、モンテカルロシミュレーションにより、Aoshima and Yata(2014)で示されている正規近似と Watanabe et al.(2019)で示した正規近似の近似精度を比較し、Watanabe et al.(2019)で示した正規近似の方が近似精度が多くの場合で良くなっていることを数値的に確認している。さらに、誤判別確率の推定量についてもシミュレーションを行い、Watanabe et al.(2019)で提案した推定量とクロスバリデーションを用いた場合の推定量について、平均二乗誤差（ＭＳＥ）を比較し、Watanabe et al.(2019)で提案した推定量の方がＭＳＥが小さくなっていることを確認している。

第５章では、第４章で述べた誤判別確率の推定問題を多群に拡張する問題について述べている。２群の場合と同様、判別関数の漸近正規性をマルチンゲール中心極限定理を用いて証明している。しかし、ここで得られる漸近近似の式には２群の場合には含まれていなかった未知パラメータも一つ含まれている。そのため、この未知パラメータの不偏推定量についても第５章で新たに導出している。さらに、モンテカルロシミュレーションにより３群の場合において、Watanabe, Hyodo and Seo(2019)で導出した正規近似の精度が Aoshima and Yata(2014)における近似精度を多くの場合で上回っていることを確認し、Watanabe, Hyodo and Seo(2019)で導出した誤判別確率の推定量のＭＳＥがクロスバリデーションを用いた場合の誤判別確率の推定量のＭＳＥよりも小さくなることを確認している。

第６章では、平均ベクトルの差のユークリッド二乗距離の同時信頼区間について述べている。第４章及び第５章では、高次元データにおけるユークリッド距離に基づく判別問題について述べたが、この判別の精度は母集団間の平均ベクトル間のユークリッド距離に依存するため、第６章の内容はユークリッド距離に基づく判別問題を考える上で有用である。平均ベクトルに関する先行研究として、例えば、Bai and Saranadasa(1996)では、２標本の平均ベクトルの検定を分散共分散行列が等しいという仮定の下で考え、ユークリッド二乗距離に基づいた推定量を導出している。また、Chen and Qin(2010)では、正規性を仮定せずに、２標本平均ベクトルの検定を導出しており、その他にも多くの平均ベクトルに関する先行研究がなされている。第６章で考える統計モデルは、第４章及び第５章で考えたモデルを含んだ、さらに適用範囲の広いモデルを考える。また、統計量についても、パラメータに適切な値を代入することで Chen and Qin(2010)で提案されている２標本の平均ベクトルの検定、Yamada and Himeno(2015)で提案されている多標本の平均ベクトルの検定、さらには線形仮説の検定等にも利用できるような抽象的な統計量を取り扱う。この統計量についてマルチンゲール中心極限定理を適用し、漸近正規性を証明している。また、

ここで示した漸近正規性は、パラメータに適当な値を代入することにより Chen and Qin(2010)で示されている漸近正規性と同様の統計量の漸近正規性が得られる。しかし、Chen and Qin(2010)では２群間のマハラノビス二乗距離にあたる項のオーダーに関する仮定を置いた下で証明している。つまり、この仮定は平均ベクトルの検定の局所対立仮説を考えていることを意味しており、この漸近正規性を用いて区間推定をするのは適さない。そこで、区間推定に応用するため、この仮定を置かない、異なる仮定の下で漸近正規性を示している。さらに、この漸近正規性を用いて平均ベクトル間のユークリッド二乗距離の区間推定を導出し、漸近的な被覆確率を与えている。さらに、モンテカルロシミュレーションにより、導出した信頼区間の経験的な被覆確率を計算している。標本サイズ及び次元数が共に小さい場合は名目上の被覆確率よりも少し小さい被覆確率をとるが、標本サイズが大きい場合、名目上の被覆確率に非常に近い値をとる傾向がみられた。仮定を満たさない多変量歪正規分布についても、近似精度がそれほど低下しない傾向も確認した。また、Chen and Qin(2010)の結果を基にした信頼区間の被覆確率は、名目上の被覆確率よりもはるかに小さい被覆確率をとることも、モンテカルロシミュレーションにより確認している。数値実験の最後には、実データへの適用例を紹介している。実データとして、Khan et al.(2001)による小児小円形青色細胞腫瘍の２３０８次元の４種類の遺伝子腫瘍のデータに対して、各遺伝子腫瘍間の平均ベクトルの同時信頼区間を構成し、ユークリッド距離のオーダーを求め、その値が大きいことを確認した。その値の大きさから、遺伝子腫瘍のような判別においては、第４章や第５章で考えたユークリッド距離による判別であっても、良い精度が得られることが期待できる。

論文審査の結果の要旨

本論文は，高次元データに対する判別分析について，主に誤判別確率の推定問題，そして関連する問題として高次元データに対する平均ベクトルの同時信頼区間について述べたものであり，６つの章から構成される．

第１章は序論である．ここでは，本論文の研究の位置づけと研究の背景について述べている．具体的には，線形判別分析と判別分析の分類精度を評価する上で重要な指標となる誤判別確率について述べている．誤判別確率の推定問題に関しては，伝統的な大標本枠組みの下で多くの結果が得られている．しかしながら，近年のビッグデーターの一つである高次元データ（次元数が標本サイズよりも非常に大きいデータ）の下では，これまでの大標本枠組みの下で確立された統計的手法を適用することはでき

ない．そこで本章では高次元データに関する先行研究を紹介するとともに，このような高次元データにも適用できる分析法について述べている．最後に本論文の各章ごとの研究内容の要旨を与えている．

第２章では，大標本枠組みにおける基本的な線形判別分析の誤判別確率に関する先行研究を詳細に紹介し，第３章では，高次元データに対する判別分析について多変量正規性の仮定や分散共分散行列が等しいという仮定を緩めた設定の下で議論している先行研究について述べている．

第４章では，高次元データの下での２群判別分析について述べている．特に，ここでは，ユークリッド距離を用いた判別関数を用いて，先行研究とは異なるある仮定の下で判別関数の漸近正規性を示すことに成功している．その証明についてはマルチンゲール差分中心極限定理を利用し，漸近正規性に基づく，誤判別確率の漸近近似を新たに導出している．さらに，誤判別確率の漸近近似の中に含まれている未知パラメータについて，その不偏推定量を与え，一致性についても証明することに成功している．なお，不偏推定量の一致性を示す際には，その分散の煩雑な計算が必要となる．最後にモンテカルロ・シミュレーションを行い，先行研究との数値比較を通して，本論文による結果の近似精度が良いことを示している．

第５章では高次元データの下での多群判別分析に対する誤判別確率の推定問題について述べている．この章で得られている結果は第４章の２群判別の場合の拡張となっている．具体的には，群の数から１を引いた個数の判別関数の同時分布が，多次元正規分布へ収束することを示し，その結果に基づいて，誤判別確率の漸近近似を与えている．また，漸近近似の中に含まれるパラメータは，第４章の２群判別にはなかった新たなパラメータが含まれており，その不偏推定量を与えることにも成功している．そして第４章と同様に，３群の場合のモンテカルロ・シミュレーションを行い，数値的評価を与えている．

第６章では，これまでの各章の結果を受けて，高次元データの下での２つの平均ベクトル間の差の同時信頼区間について述べている．ここでの同時信頼区間は平均ベクトル間のユークリッド距離を基に構成され，特筆すべき点は，考える統計量として，平均ベクトルの検定ばかりでなく，線形仮説の検定にも適用できるように拡張した統計量を取り扱い，その漸近正規性を示すことに成功しているところである．最後にモンテカルロ・シミュレーションを行い，導出した信頼区間の数値的な被覆確率と名目上の信頼度を比較することにより評価を与えている．そして高次元データである実データに適用した解析例も与えている．

以上により，これらの結果は多変量統計解析の理論に対する大きな貢献である．よって，本論文は博士（理学）の学位論文として十分価値があると認める．