Statistical hypothesis test for some coefficients of the linear discriminant function with two-step monotone missing data

岡田, 悠希

(Degree)
博士（理学）

(Date of Degree)
2025-03-25

(Resource Type)
doctoral thesis

(Report Number)
甲第9138号

(URL)
https://hdl.handle.net/20.500.14094/0100496419

（別紙様式 3）

論 文 内 容 の 要 旨

氏　　名　　　岡田　悠希

専　　攻　　　理学研究科数学専攻

論文題目（外国語の場合は，その和訳を併記すること。）

Statistical hypothesis test for some coefficients of the linear discriminant function with two-step monotone missing data
(2-ステップ単調欠測データで構成される線形判別関数の係数に対する統計的仮説検定)

指導教員　　　首藤　信通

多変量解析において，変数選択問題は統計解析に必要となる変数を抽出する重要な問題として知られている (例えば [1] を参照). 特に測定に侵襲を伴うことが多い医療・看護分野では十分な標本サイズを確保することが困難である上，観測対象が研究期間中に脱落することで欠測データとなる状況が多く見られる．この問題に対して本博士論文では，2-step 単調欠測データに基づく線形判別関数における一部の係数に関する冗長性検定を尤度比規準によって与え，導出した検定統計量に対する漸近的なバイアス修正 (Bartlett 修正) による修正型検定を構成する ([3]).

p 次元正規母集団 $\Pi^{(g)} : N_p(\boldsymbol{\mu}^{(g)}, \Sigma)$ $(g \in \{1,2\})$ から得られた 2-step 単調欠測データを

$$\begin{pmatrix} \boldsymbol{x}_{11}^{(g)} & \boldsymbol{x}_{12}^{(g)} & \cdots & \boldsymbol{x}_{1N_1^{(g)}}^{(g)} & \boldsymbol{x}_{1N_1^{(g)}+1}^{(g)} & \cdots & \boldsymbol{x}_{1N^{(g)}}^{(g)} \\ \boldsymbol{x}_{21}^{(g)} & \boldsymbol{x}_{22}^{(g)} & \cdots & \boldsymbol{x}_{2N_1^{(g)}}^{(g)} & \boldsymbol{x}_{2N_1^{(g)}+1}^{(g)} & \cdots & \boldsymbol{x}_{2N^{(g)}}^{(g)} \\ \boldsymbol{x}_{31}^{(g)} & \boldsymbol{x}_{32}^{(g)} & \cdots & \boldsymbol{x}_{3N_1^{(g)}}^{(g)} & * & \cdots & * \end{pmatrix},$$

$$\boldsymbol{x}_{(12)j}^{(g)} = \left(\boldsymbol{x}_{1j}^{(g)\prime}\ \boldsymbol{x}_{2j}^{(g)\prime}\right)',\ \boldsymbol{x}_{1j}^{(g)} = \left(x_{1j}^{(g)}\ \cdots x_{p_1 j}^{(g)}\right)',$$

$$\boldsymbol{x}_{2j}^{(g)} = \left(x_{p_1+1j}^{(g)}\ \cdots\ x_{p_1+p_2 j}^{(g)}\right)',\ \boldsymbol{x}_{3j}^{(g)} = \left(x_{p_1+p_2+1j}^{(g)}\ \cdots\ x_{pj}^{(g)}\right)'$$

とする．ここで，$j \in \{1, 2, \cdots, N^{(g)}\}$, $N^{(g)} > N_1^{(g)}$, $p_3 = p - p_1 - p_2$ である．また，“$*$” は欠測を示す．

一方，判別すべき p 次元標本ベクトル $\boldsymbol{x} = (\boldsymbol{x}_1'\ \boldsymbol{x}_2'\ \boldsymbol{x}_3')'$ に対するパラメータ $\boldsymbol{\mu}^{(g)}$, Σ が既知の場合の線形判別関数 W_0 は

$$W_0 = (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})'\Sigma^{-1}\left\{\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)})\right\}$$

であるが，$\boldsymbol{\eta}_2^{(1)} = \boldsymbol{\eta}_2^{(2)}$, $\boldsymbol{\eta}_3^{(1)} = \boldsymbol{\eta}_3^{(2)}$ の下で

$$W_0 = (\boldsymbol{\mu}_1^{(1)} - \boldsymbol{\mu}_1^{(2)})'\Sigma_{11}^{-1}\left\{\boldsymbol{x}_1 - \frac{1}{2}(\boldsymbol{\mu}_1^{(1)} + \boldsymbol{\mu}_1^{(2)})\right\}$$

となる．ただし $\boldsymbol{x}_i = C_i\boldsymbol{x}$, $\boldsymbol{\mu}_i^{(g)} = C_i\boldsymbol{\mu}^{(g)}$ $(i \in \{1,2,3\})$, $\Sigma_{ik} = C_i\Sigma C_k'$ $((i,k) \in \{1,2,3\}^2)$,

$$C_1 = (I_{p_1}\ O_{12}\ O_{13}),\ C_2 = (O_{21}\ I_{p_2}\ O_{23}),\ C_3 = (O_{31}\ O_{32}\ I_{p_3}),$$

$$\boldsymbol{\eta}_2^{(g)} = \boldsymbol{\mu}_2^{(g)} - \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\mu}_1^{(g)},\ \boldsymbol{\eta}_3^{(g)} = \boldsymbol{\mu}_3^{(g)} - \Sigma_{3(12)}\Sigma_{(12)}^{-1}\boldsymbol{\mu}_{(12)}^{(g)},$$

$$\boldsymbol{\mu}_{(12)}^{(g)} = \begin{pmatrix} \boldsymbol{\mu}_1^{(g)} \\ \boldsymbol{\mu}_2^{(g)} \end{pmatrix},\ \Sigma_{(12)} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},\ \Sigma_{3(12)} = \Sigma_{(12)3}' = (\Sigma_{31}\ \Sigma_{32}),$$

I_{p_i} は p_i 次元単位行列, O_{ik} は $p_i \times p_k$ 次元零行列である．よって，$\boldsymbol{x}_2$ および $\boldsymbol{x}_3$ の係数が冗長であるという帰無仮説を検定する仮説検定問題は

$$H_0 : \boldsymbol{\eta}_2^{(1)} = \boldsymbol{\eta}_2^{(2)},\ \boldsymbol{\eta}_3^{(1)} = \boldsymbol{\eta}_3^{(2)} \quad \text{vs.} \quad H_1 :\neq H_0$$

に対する仮説検定問題に帰着する．このことより，第 2 章では，[2] で示された密度の分解を用いて 2-step 単調欠測データにおける帰無仮説 H_0 の下での最尤推定量 $\left(\widetilde{\boldsymbol{\mu}}^{(1)},\ \widetilde{\boldsymbol{\mu}}^{(2)},\ \widetilde{\Sigma}\right)$，および対立仮説 H_1 の下での最尤推定量 $\left(\widehat{\boldsymbol{\mu}}^{(1)},\ \widehat{\boldsymbol{\mu}}^{(2)},\ \widehat{\Sigma}\right)$ をそれぞれ導出し，尤度比検定統計量 $-2\log\lambda$ を与えた．ここに，λ は 2-step 単調欠測データによって構成される上記の仮説検定問題に対する尤度比である．なお，$-2\log\lambda$ は 2-step 単調欠測データにおける大標本漸近枠組み ([4]) および H_0 の下で漸近的に自由度 p_2+p_3 のカイ二乗分布に従う．第 3 章では，H_0 および大標本漸近枠組みの下で第 2 章で導出した $-2\log\lambda$ の期待値に対する漸近展開を求めることで，Bartlett 修正項を導出し，それを用いて与えられる修正型尤度比検定統計量 $-2\log\lambda^*$ を提案した．ここに，λ^* は修正後の尤度比である．修正項は未知のパラメータを含む Mahalanobis 二乗距離 $\Delta_{11}^2 = \left(\boldsymbol{\mu}_1^{(1)}-\boldsymbol{\mu}_1^{(2)}\right)'\Sigma_{11}^{-1}\left(\boldsymbol{\mu}_1^{(1)}-\boldsymbol{\mu}_1^{(2)}\right)$ に依るため，実用上は $-2\log\lambda^*$ に [4] で提案された Δ_{11}^2 の漸近不偏推定量を代入して得られる検定統計量 $-2\log\widehat{\lambda}^*$ を用いることとなる．第 4 章では，第 2 章において導出した $-2\log\lambda$ と第 3 章で提案した修正型検定統計量 $-2\log\widehat{\lambda}^*$ に対してカイ二乗分布を帰無分布の近似として用いた場合の仮説検定方式について数値的な評価を与えた．特に，帰無仮説 H_0 の下で，さまざまなサンプルサイズの設定の下で反復回数 100,000 回のモンテカルロ・シミュレーションを実行し，検定のサイズ (帰無仮説の下で帰無仮説を誤って棄却する確率) を比較した結果，実行したすべてのケースで $-2\log\widehat{\lambda}^*$ による仮説検定方式の方が検定のサイズが有意水準との差が小さいことが確認された．第 5 章では，本博士論文のまとめと，関連する今後の問題を挙げた．付録として第 2 章において最尤推定量の導出 (Theorem 2.1, 2.2) に用いた補題と，第 3 章における主結果 (Theorem 3.1) の導出に用いた期待値の証明を付している．

参考文献

[1] M. Hyodo and T. Kubokawa (2014). A variable selection criterion for linear discriminant rule and its optimality in high dimensional and large sample data, *Journal of Multivariate Analysis*, **123**, 364–379.

[2] T. Kanda and Y. Fujikoshi (1998). Some basic properties of the MLE's for a multivariate normal distribution with monotone missing data, *American Journal of Mathematical and Management Sciences*, **18**, 161–190.

[3] Y. Okada, N. Matsuchi, M. Hyodo, N. Shutoh (in press). Redundancy test for some variables in linear discriminant analysis with two-step monotone incomplete data, *SUT Journal of Mathematics*.

[4] N. Shutoh, M. Hyodo and T. Seo (2011). An asymptotic approximation for EPMC in linear discriminant analysis based on two-step monotone missing samples, *Journal of Multivariate Analysis*, **102**, 252–263.

（別紙１）　　　　　　　　　　　　　論文審査の結果の要旨

氏名	岡田　悠希		
論文題目	Statistical hypothesis test for some coefficients of the linear discriminant function with two-step monotone missing data （2-ステップ単調欠測データで構成される線形判別関数の係数に対する統計的仮説検定）		
審査委員	区　分	職　名	氏　　　名
	主　査	教　授	首藤　信通
	副　査	教　授	青木　敏
	副　査	教　授	福山　克司
	副　査	教　授	谷口　隆
	副　査		

要　　旨

本論文は，母集団分布が多変量正規分布である２個の群（母集団）から，それぞれ2-ステップ単調欠測データ（欠測パターン数２の単調欠測データ）が観測される仮定の下で，線形判別分析における変数選択問題を扱っている。特に，本論文はすべての標本ベクトルで観測されている変数の一部と一部の標本ベクトルで欠測している変数のすべてが線形判別分析において冗長であるという帰無仮説，すなわち，これらの変数に対応する線形判別関数の係数が零であることを帰無仮説とする統計的仮説検定を構成し，その修正法を与えるものである。

本論文の構成は次の通りである。第１章は序論である。2-ステップ単調欠測データを仮定した統計的推測法の先行研究についてまとめており，本論文の研究の位置づけをおこなうとともに，線形判別分析において冗長性検定が線形判別関数の係数に対する仮説検定問題に帰着することについて述べている。第２章では帰無仮説，対立仮説それぞれの下で得られる最尤推定量を導出し，尤度比検定統計量の構成をおこなっている。第３章では主結果を導出している。具体的には，第２章で得られた尤度比検定統計量の漸近期待値を求め，その結果を用いて尤度比検定統計量に対するBartlett修正（検定統計量に対する漸近バイアスの修正）を与えることによって得られるBartlett修正検定を提案している。第４章ではさまざまな平均ベクトルや標本サイズの設定の下でモンテカルロ・シミュレーションをおこない，第２章で構成した尤度比検定と第３章で提案したBartlett修正検定のそれぞれに対し，検定のサイズ（帰無仮説が成り立つ下で帰無仮説を棄却する確率）を数値的に評価している。その結果，数値実験を実行したすべての設定の下で，第３章で提案したBartlett修正検定のサイズの方が，第２章で構成した尤度比検定のサイズよりも有意水準との差が小さい値となっており，その意味でBartlett修正検定の方が正確な仮説検定であることが数値的に確認されている。第５章は本論文の結論を与えている。付録では，第２章，第３章の理論的結果の導出に必要となる補題を示している。

氏名	岡田　悠希

本論文の研究テーマである欠測データ解析は主に 1960 年代後半から理論研究が進められており，欠損値に対する代入法，最尤推定量の性質，直接尤度に基づく多変量解析手法に関する研究結果が多く得られている。特に，観測対象の脱落によって生じる単調欠測データに対する多変量解析手法に関する理論研究は，最尤推定量の分布特性（Kanda & Fujikoshi (1998)）が得られたことを契機として，その後の発展が著しい状況である。しかしながら，近年，標本ベクトルの次元数が標本サイズを超えるデータを扱う高次元データ解析などでも研究対象として扱われるようになった変数選択問題は，欠測データの下では最尤推定量の表現の複雑さから十分に議論がなされておらず，理論・応用の両面においてその研究の進展が期待されている状況にある。したがって，岡田悠希氏が本論文で扱う研究課題は，欠測データ解析の理論研究において挑戦的な取り組みであると判断される。また，第 2 章では本論文で扱う尤度比検定統計量の構成をおこなっているが，第 3 章の主結果は尤度比検定統計量が得られた形式のまま求めようとすると計算が非常に煩雑になり，その導出が困難になることが予想される。岡田悠希氏は尤度比検定統計量の表現を可能な限り簡略化した後に第 3 章の主結果を導出しており，その点においても研究上の工夫がみられる。一般に，平均ベクトルと共分散行列を含む帰無仮説に対する検定統計量の Bartlett 修正を求めることは難しく，本論文の主結果の導出には Wishart 行列の 4 次モーメントの結果の一般化が必要となる。岡田悠希氏らの研究論文がその証明を与えており，この新たな理論的結果によって他の仮説検定問題の検定統計量に対する Bartlett 修正を与えることが可能となる等の波及も予想される。

本研究は，欠測データに基づく線形判別分析について，その変数選択に関する仮説検定問題を研究したものであり，上記の創意工夫とともに尤度比検定統計量に対する Bartlett 修正検定について重要な知見を得たものとして価値ある集積であると認める。よって，学位申請者の岡田 悠希は，博士（理学）の学位を得る資格があると認める。