R. A. Fisher 以後の判別分析の新理論と遺伝子解析の新手法

成蹊大学　経済学部　特任教授　新村秀一

Fisher は「Fisher の仮説」を考えることで、計算機の助けを借りないで判別理論を確立した。しかし、Fisher の仮説を検定する良い統計量がないこと、誤分類確率や判別係数の標準誤差が定式化できなかったこと、などの重要な事実に多くの研究者や利用者は注意を払わなかった。Fisher も指摘する通り、「Fisher の仮説を満たさないデータに Fisher の LDF を適用してはいけない」が、それが分からないため多くの分野に適用され、一見成果を上げてきた。しかし大学卒業後の 1971 年に「心電図の自動解析システムの診断論理を LDF と QDF で 4 年間アプローチして、医師の開発した枝分かれ論理に歯が立たなかった」。この経験から、医学診断や、各種の格付けや、試験の合否判定を得点合計で判別すると、線形分離可能（MNM=0）であっても誤分類確率が最高 20%になることが実証研究で分かった。即ち既存の判別理論には、世界中の誰も研究していない 4 つの深刻な瑕疵があることを判別分析の多くの実証研究で見つけ、数理計画法による 5 個の最適判別関数を開発し解決してきた。また、判別分析は推測統計手法でないので、「小標本のための 100 重交差検証法（新手法 1）」を開発し、 誤分類確率と判別係数の 95%信頼区間を求めることができた。また、検証標本の平均誤分類確率（M2）が最小モデルを Best モデルとして選ぶと、Vapnik が提起した汎化能力に優れたモデルを選ぶことになる。各手法の全ての変数の組み合わせモデルで Best モデルを選び、8 種の異なった LDF（改定 IP-OLDF、改定 LP-OLDF、改定 IPLP-OLDF、H-SVM、SVM4、SVM1、ロジスティック回帰、Fisher の LDF）の Best モデルの中で最小の M2 を比較して、多くの場合で改定 IP－OLDF の Best モデルの M2 が最小であった。Vapnik の定義した汎化能力は、単に固定された p 変数の LDF での話であることが理解されていない。また、新手法 1 は、LOO 法よりはるかに優れている。以上で 4 種の深刻な問題を全て解決したが、2015 年の富山での科研費シンポジュームで「世界的に著名な Microarray データが公開されているのを知り、それらを判別すると、世界中の統計や医学研究者が高次元空間の分析と称し、特に Feature Selection の研究を 15 年以上行ってきたが(問題 5)が、改訂 IP-OLDF で解決した（Matroska Feature Selection Method、新手法 2)。そして Microarray データは、MNM=0 になる小さな部分空間の排他的な和集合になっていることが分かった」。即ち問題 5 の認識と同時に、僅か 41 日間で新手法 2 を開発した。これによって、新手法 1 と新手法 2 で Fisher 以後の新しい判別理論を確立した（New Theory of Discriminant Analysis After R. Fisher, Springer(2017)）。

参加者と以下の点について真摯に議論したい.

1) Fisher は Fisher の仮説に基づいて判別分析理論を確立したが，限界を知っていた．田辺の指摘[48]の他，検証にアヤメのデータを用いている点，仮説を満たさない場合に QDF が提案されていることが重要である．
2) Fisher 以後，RDA や LASSO 等の分散共分散に基づく理論が開発され，これらが Fisher の後継者と考えられているが間違いである．Fisher 理論を適用してはいけない医学分野で，Cox は Cox 回帰やロジスティック回帰を提案したが，彼こそが Fisher の正当な後継者である．そして Vapnik は MP で判別分析にアプローチした。統計や OR の分野を避け、多くの実証研究をパターン認識の分野で汎化能力や LSD 判別やカーネル SVM を広めた第三世代の後継者である．
3) 新村は，既存の判別理論には 4 つの問題があることを実証研究で見つけ，それらを解決した．その応用研究として，統計や医学の多くの研究者が従来の統計手法で 10 年から 15 年以上に渡り高次元空間の分析を試みたが成果は得られなかった．しかし 3 種類の OLDF は高次元空間を少数の遺伝子空間に自然に縮約し，Matroska 構造をもつこと，Datasets は少数の排他的な SM の和集合であることを示した．すなわち，これらの各 SM は統計的に小標本であり，通常の統計手法で分析可能である．
4) LASSO 等のアプローチは無意味である．通常のスイス銀行や日本車データで LSD であることが分からず、Feature Selection をできないのに高次元でできると考えるのは論理的ではない．