様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　５　月　３１　日現在

機関番号：　１３６０１
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６７８０１３７

研究課題名（和文）識別不可能な有限混合モデルの推定と要素密度の個数特定化のための検定統計量の開発

研究課題名（英文）Estimation of unidentifiable finite mixture binary models and test statistics for specifying the number of components

研究代表者

増原　宏明（MASUHARA, Hiroaki）

信州大学・学術研究院社会科学系・准教授

研究者番号：１０４１９１５３

交付決定額（研究期間全体）：（直接経費）　　2,000,000円

研究成果の概要（和文）：本課題では、2つ以上の分布が加法的に混ざっている有限混合モデルにおいて未解決な2つの問題を研究した。1つ目が、0と1の値をとる2値変数においては有限混合モデルが識別不可能である問題であり、正規分布に従う連続変数との同時方程式モデルであれば識別可能となることを証明した。2つ目が、有限混合モデルでは何個の分布から成り立っているかを検定できない問題である。過剰識別のために尤度が際限なく上昇するので、ラプラス近似を用いたVuong検定を行い、個数の特定化のためのシミュレーションを行った。その結果、限定的ながらもラプラス近似が望ましいことが確認できた。

研究成果の概要（英文）：We analyze two problems in a finite mixture model in which two or more distributions are additively mixed. First, the finite mixture model can not be distinguished in a binary variable taking values of 0 or 1. We demonstrate that it can be identified in a simultaneous equation model with a continuous variable following the normal distribution. Second, the finite mixture models can not be tested how many distributions it is composed of. Since the likelihood increases endlessly due to over-identification, we propose a Vuong test using Laplace approximation. In Monte-Carlo simulation, Laplace approximation is preferable.

研究分野：　応用計量経済学

キーワード：　有限混合モデル　識別性　検定統計量

１．研究開始当初の背景

有限混合（finite mixture, FM）モデルとは、ある確率（被説明）変数が単一の確率密度関数から生成されるのではなく、複数ある要素密度の中の１つから生成されるが、どの要素分布から生成されたかは観測されないような場合に用いられる統計モデルである。応用経済学や他の社会科学分野で用いられるミクロデータは、実験計画法によって完全に制御されたデータではなく、観測不可能な異質性を含むことを排除できない。この観測不可能な異質性を確率変数として捉える FM モデルは、頑健な推定結果が期待できるだけでなく、集団を「タイプ」として表すことができ、経済学的な解釈の幅が広がることが期待される。

ところが実証分析において FM モデルは頻繁に利用されない。それは以下に示す 2 つの問題を解決できないためである。第 1 の問題は被説明変数によっては識別不可能となることであり、第 2 の問題は要素の個数が何個であるかを検定によって決定することができない点である。本研究は上記の 2 つの問題を解決するために実施した。

２．研究の目的

本研究の目的は、先に述べた有限混合の 2 つの未解決の問題を分析することである。これは以下のようにまとめられる。

（1）有限混合モデルにおいて、被説明変数が 2 値（0 と 1）の横断面データの場合に、識別不可能となり、推定できない

（2）有限混合モデルにおいて要素密度の個数を尤度比検定で特定化しようとも、これが適用できない

上記の（1）については、大規模ミクロデータを用いて分析するときに頻繁に遭遇する。パネルデータではなく横断面データでの分析であると、2 値の被説明変数の場合には、FM モデルを適用できないことになる。結果的に、伝統的に使用されているプロビット・モデルもしくはロジット・モデルしか使用できない。しかしながら、非実験的環境で得られたデータは、観察不可能な異質性を排除できず、2 値の横断面データでも有限混合分布を適用するほうが望ましい場合もある。そこで本研究では、識別可能な代替的な方法の開発を試みた。

上記の（2）の問題は、FM モデルを使用する場合に常に発生する。これは過剰識別によって尤度が際限なく上昇するためである。そのため、伝統的に使用されている尤度比検定を用いることができず、先行研究の多くでは検定統計量ではない情報量基準に頼って要素密度の個数の特定化を行っていた。しかし検定により特定化できれば、より科学的に定式化ができるものと期待される。そこで本研究では非入れ子型の検定統計量を用い、尤度を様々な方法で近似し、代替的な特定化方法を検証した。

３．研究の方法

（1）識別不可能な有限混合モデル

識別不可能性は 2 値の横断面データの FM モデルにおいて生じ、識別不可能性を回避するためには、新たに代替案を構築しなければならない。大規模横断面ミクロデータにおいて、被説明変数が 2 値データのみとは考えにくく、複数の多値変数や連続変数が存在するのが普通である。そこで、多変量の 2 値データや、2 値データと相関をもつ連続データなどの同時方程式モデルを考え、これらのモデルで識別不可能性が生じるのかを証明した。証明方法は先行研究を拡張したものを用い、積率母関数で識別可能かどうかを検証した。識別可能であると証明がなされたら、モンテカルロ・シミュレーションを試み、現実的なデータ発生過程においても識別可能かどうか、そのパフォーマンスを分析した。

（2）有限混合モデルの要素の個数特定化について

要素密度の個数を決定するのに尤度比検定では過剰識別による境界問題から使用できない。また、FM モデルは過剰識別が生じると、尤度が際限なく増加することがあるので、真の尤度を何らかの方法で近似しなければならない。本研究では、以下の 2 つのステップを踏んで、検定統計量の開発を行った。

1. 尤度の近似方法を複数提案する
2. 非入れ子型のモデル選択に使用できる Vuong 検定が適用可能かどうか、近似された尤度で計算した検定統計量の極限分布を調べる

上記 1 の尤度の近似には、3 つの方法を試した。第 1 の方法がラプラス近似、第 2 の方法が識別尤度（classification likelihood）、第 3 の方法が識別尤度をラプラス近似したものである。

４．研究成果

（1）識別不可能な有限混合モデル

3 つのステップを踏んで証明を行った。第 1 に、先行研究と同様の積率母関数を用いる証明により、2 変数の有限混合モデルの識別可能性に関する補題を証明した。この補題では、2 つの変数のうちいずれかの変数の性質が、識別可能性に影響を及ぼすことが示された。第 2 に、この補題を用いて、2 値変数同士の同時方程式モデルを検証した。この場合は、2 値変数が 1 つのときと同様に、識別不可能であることが示された。第 3 に、第 1 で述べた補題を用いて、2 値変数と連続変数の

同時方程式モデルを検証した。積率母関数に代数計算できない積分が含まれたので、ガウス型の積分公式を用いて、これを近似した。その結果、識別可能となることが示された。またこの場合に、連続変数は2値変数と相関があっても無相関であっても、識別性に影響が無いことも明らかとなった。

上記の証明結果を、モンテカルロ・シミュレーションで確認したところ、連続変数に正規分布を置き、2 個の要素密度において、平均値パラメータを同一に設定し、標準偏差パラメータのみを変化させるという比較的厳しい条件においても、2 値の有限混合モデルは識別可能であることが示された。すなわち、2 値変数が単独で存在すると、有限混合モデルを用いることはできないが、これとは別に連続変数が1つでも存在すれば、有限混合モデルによる推定が可能となることが示された。この結果は、実際のデータで推定する場合には意味を有する。なぜならば、意思決定を表す2値の変数以外に、所得などの内生変数が存在すれば、2 つの同時方程式を推定することで、識別不可能であった2値の変数は FM モデルとして推定できることを意味するからである。

（2）有限混合モデルの要素の個数特定化について

有限混合モデルにおいては、有限の要素密度で推定するので、要素密度を足すことでいくらでも密度の近似の程度が上がり、過剰識別が生じやすい。これにより、尤度が際限なく増加するために、モデル選択において頻繁に使用する尤度比検定を適用することができない。

そこで、この問題を回避するための方法を実施した。具体的には、Vuong の非入れ子型の検定統計量を用いて、要素密度の個数特定化を試みた。そこでは当然のことながら、先に議論した過剰識別を回避する必要があった。そこで、推定された尤度が正しくないという前提の下、尤度を近似する複数の方法を試した。具体的には、ラプラス近似、識別尤度（classification likelihood）、これら 2 つの折衷案の統合識別尤度である。

まずラプラス近似とはベイズの情報量基準の核となる理論で、分散共分散行列の推定量である情報行列を用いて尤度を補正しようという試みで、正しい定式化の場合には標準偏差が小さくそのペナルティも小さくなるが、誤った定式化の場合はペナルティが大きくなることを利用している。次に識別尤度とは、尤度を期待完全尤度のみで評価する方法である。真の尤度で、要素密度の個数が増加するにつれて罰則を受けることを利用し、この識別できない部分を差し引いて、識別できる部分のみを用いることに特徴がある。

モンテカルロ・シミュレーションの結果、以下のような結論を得ることができた。通常の尤度では、尤度が過大になるため有限混合モデルの個数が大きいものを採択しやすく、逆に識別尤度、統合識別尤度では尤度を過少に評価することから、有限混合モデルの個数が小さいものを採択しやすいことが認められた。すなわち、通常の尤度、識別尤度、統合識別尤度からは、有限混合モデルの個数を決定することはできなかった。しかしながら、ラプラス近似を用いた Vuong 検定であれば、真の定式化が未知であっても、サンプルサイズが大きくなれば検出力が高くなり、第1種過誤も小さい傾向を有していた。すなわち、ラプラス近似を用いた検定であれば、真の定式化を統計的に決定できる可能性があると結論付けることができた。

最後に、上記の（1）と（2）で得られた研究成果を用いて、現実のデータに応用する研究を行った。データは、日本版総合社会調査（Japanese General Social Survey、JGSS）や老研－ミシガン大全国高齢者パネル調査等を予定していたが、当該分野での研究成果が蓄積されておらず、先行研究との比較が誰でもわかるようにするために、より汎用性の高いデータを用いることとした。具体的には British Household Panel Survey (BHPS)、および Health and Lifestyle Survey (HALS) である。これらのデータの一部は教育目的で完全に公開されており、また多くの研究者にとってはなじみのあるデータであるので、これを採用した。研究結果を現在まとめている段階であり、順次投稿する予定である。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　３件）

①櫻井秀彦、丹野忠普、増原宏明、林行成、恩田光子、山田玲良、医療用医薬品の流通分析―卸の機能と情報提供サービスに関する実証研究―、流通研究、査読有、Vol.19、No.1、2016、pp.1-10

②増原宏明、有限混合モデルの個数特定化のための検定統計量に関する一考察、広島国際大学医療経営学論叢、査読有、Vol.8、2016、pp.61-76

③増原宏明、小西幹彦、丁井雅美、林行成、保険者医療費データによる生涯医療費シミュレーションのための統計理論、日本医療経営学会誌、査読有、Vol.9、No.1、2016、pp.47-56

〔図書〕（計　１件）

①MASUHARA Hiroaki, 一橋大学学位請求論文博士（経済学）（一橋大学・経第 178 号）, Essays on Unobserved Heterogeneity and Endogeneity in Health Econometrics (医療計量経済学における観察不可能な異質性と内生性に関する諸研究), 2014, 112

〔産業財産権〕

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
増原　宏明（MASUHARA, Hiroaki）
信州大学・学術研究院社会科学系・准教授
研究者番号：１０４１９１５３

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）