九州大学学術情報リポジトリ
Kyushu University Institutional Repository

切断分布をともなう回帰分析におけるKernel-based手法の適用 : サンプル間変数共有の分析を中心として

時永, 祥三
九州大学大学院経済学研究院

譚, 康融
久留米大学経済学部 : 教授

https://doi.org/10.15017/18622

出版情報：經濟學研究. 77 (2/3), pp.35-44, 2010-09-30. Society of Political Economy, Kyushu University
バージョン：
権利関係：

切断分布をともなう回帰分析における Kernel-based 手法の適用

── サンプル間変数共有の分析を中心として ──

時　永　祥　三
譚　　康　融

1　まえがき

多変量解析など統計解析において, 変数を高次元の変数へと写像 (非線形変換) することを用いて推定の精度を向上させる方法が提案されており, Kernel-Based(KB) 手法は代表的な方法である [1]-[5]。これまで多変量解析の分野においては, 回帰分析, 判別分析や主成分分析へと KB 手法を適用することがなされ, その有効性が示されている [1]-[5]。一方, 変数の分布が切断分布 (被説明変数が観測されないで, 通常は処理から除外されるサンプルを含む分布) にしたがう場合は, 社会調査などではしばしば観測される事象であり, 回帰分析による解析においても, 最小 2 乗法などの通常手法とはやや異なる関数最適化が必要となる [6]-[12]。しかしながら, KB 手法を切断分布をともなう回帰分析へ応用した例は示されていない。本論文では, 切断分布をともなう回帰分析における KB 手法の適用を示す [13]。この場合,KB 手法により得られる尤度が従来の回帰分析手法の水準より高くなることを, サンプル間変数共有の現象として説明する。

本論文ではまず, KB 手法の基本手順を整理し入力変数に対する非線形変換関数を適用し, この変数 (KB 変数と呼ぶ) の線形判別関数を用いた問題へと帰着させることを述べる。同時に切断分布を含む場合の回帰分析の概要についてまとめる。この中で変数が切断分布を含むケースにおいては最大化すべき尤度の中には, 求めるべきパラメータを含んだ正規分布関数と分布関数から構成される関数を最適化する問題に帰着できることを述べる。この場合, 通常の線形および非線形の回帰分析において残差から計算される最大尤度が決められるが, KB 手法適用の大きな特徴として, この水準を超える尤度が得られることをサンプル間の変数 (KB 変数) 共有の現象として説明する。同時に KB 手法においては, 一般に変数の個数がサンプル数に応じて増加する問題があるので, この情報共有の効果を変数個数を制限する目的に利用する方法を提案する。応用例として, シミュレーションにより人工的に生成されたデータおよび消費者ローンのデータへ本論文の手法した場合の性能評価と, 変数個数の制約の有効性を議論する。

以下では,2. においては KB 手法による回帰分析と切断分布について述べる。3. では,KB 手法により得られる尤度の性質を分析し, これに基づく変数個数の制限について議論する。4. においては応用例について述べる。

2 KB手法による回帰分析と切断分布

2.1 KB手法の基本

以下では, 本論文で用いるKB手法について概要を整理する [1]-[5]。なおここで考察する問題は, 基本的には回帰分析であり, 判別分析や主成分分析と比較すると分かりやすい議論であるので, 必要最小限の範囲の記述にとどめる。

いま回帰分析における被説明変数を y, 説明変数ベクトルを $x = (x_1, x_2, ..., x_n)$ としておく。またサンプルの全体を通じて変数 y, x に対応する k 番目のサンプル値を, y_k および $x_k = (x_{1k}, x_{2k}, ..., x_{nk}), k = 1, 2, ..., M$ とする。回帰分析の問題は, 変数ベクトルの M 個のサンプルを用いて, 次のような関係式の係数 $\alpha_i, i = 1 \sim n$ を推定する問題であり, 推定誤差の2乗を最小化する最小2乗法が代表的な手法である。

$$y_k = Cx_k^T + \varepsilon_k, C = (\alpha_1, \alpha_2, ..., \alpha_n), x_k = (x_{1k}, x_{2k}, ..., x_{nk}), \tag{1}$$

ここで ε_k は k 番目のサンプルについての残差である。KB手法おいては, 変数 $X = (x_1, x_2, ..., x_n)$ を非線形変換写像 $\phi(X)$ を用いて高次元の関数に変換する。これにより, より自由度の高い関数の構成が可能となる。KB手法における回帰分析は, この変換された変数の線形の回帰係数を推定する問題である。これを形式的に記述すると, 次のようになる。

$$X \to \phi(X) \tag{2}$$

$$y = \sum_{i=1}^{m} a_i \phi(x_i) \tag{3}$$

しかしながらKB手法おいては, 表現定理により, 通常の線形回帰分析の回帰係数が変数 α_i についての係数となっている表現ではなく, サンプルに関する係数 a_i による表現に変換できることに大きな違いがある [5]。すなわち, 次のような表現になる。

$$y = \sum_{j=1}^{n} a_i K(X, X_j) \tag{4}$$

ここで, $K(X, X_j)$ は, いわゆるカーネル演算であり, この計算の過程で, 写像 $\phi(X)$ による計算を簡便化するための表現を用いる。例えば2次元ベクトル (a_1, a_2) を3次元 $(a_1^2, \sqrt{2}a_1a_2, a_2^2)$ という2次モーメントへ変換する変換がある。このとき,2つのベクトル $(a_1^2, \sqrt{2}a_1a_2, a_2^2)$,$(b_1^2, \sqrt{2}b_1b_2, b_2^2)$ の要素の, すべての組み合わせの積和 $\sum_{i1,i2,j1,j2=1}^{2} a_{i1}a_{i2}b_{j1}b_{j2}$ を計算することを dot product とよぶ。これを $g(a, b) = (a \cdot b)^2$ として表現する。また, $g(a, b) = \exp(-||a - b||/\sigma)$ の形の計算をするものはガウシアン (Gaussian)Kernel と呼ばれる。本論文ではこのガウシアン Kernel を用いる。2つのサンプル i, j の変数ベクトル x_i, x_j から計算される次の変数を, KB変数と呼んでおく。

$$s_{ij} = \exp(-||x_i - x_j||/\sigma) \tag{5}$$

2.2 切断分布における回帰:トービット分析

通常の回帰分析では, 被説明変数は連続的な値をとることが仮定されているので, モデルを推定する場合には, 特に変数のとる値についての制約を考慮する必要はない。しかし消費者の行動など

をモデル化する場合には, 制約が必要となる。例えば消費者が車を購入する場合に, 収入 x と購入価格 y との関係をモデルで記述するには, 購入しない消費者の購入価格 y はゼロとする制約が加わる。このような変数の分布を切断分布と呼んでいる [6]-[12]。一般的には, 購入しなかった消費者の情報は利用されないが, 購入しないケースも含めて分析することができれば, 精度が向上することが期待できる。このような前提のもとで定式化されたものとして, トービット分析 (Tobit model), およびこれを拡張したサンプルセレクション分析 (Sample Selection model) がある。

トービット分析は, 通常の回帰分析とは異なり, 被説明変数 y_k が, ある条件を満足した場合にだけに値をもつモデルであり, 次のように書くことができる。

$$y_k^* = Ax_k^T + \varepsilon_k, A = (a_1, a_2, ..., a_n), x_k = (x_{1k}, x_{2k}, ..., x_{nk}), k = 1, 2, ..., M \tag{6}$$

$$y_k = \begin{cases} y_i^* & \text{if } y_k^* > 0; \\ 0 & \text{if } y_k^* \leq 0 \end{cases} \tag{7}$$

ここで残差 ε_k は平均ゼロ, 分散が σ^2 である正規乱数にしたがうと仮定する。なお y_k^* が負である場合にはその数値は観測されないで, 符号だけが観測される。いま, $\phi(.), \Phi(.)$ を, それぞれ, 標準正規分布の密度関数および分布関数としておく。$y_k = 0$ となる確率は, 次のように計算される。

$$Prob(y_i = 0) = Prob(Ax^T + \varepsilon_k \leq 0) = Prob(\varepsilon_i \leq -Ax^T) = \Phi(-Ax^T/\sigma) \tag{8}$$

同様に, $y_i^* > 0, y_i = y_i^*$ となる確率 (尤度) は, 次のように計算される。

$$f(y_i = y_i^* | y_i^* > 0) = Prob(y_i^* > 0) = f(y_i = y_i^*) = \Phi[(y_i - Ax^T)/\sigma] \tag{9}$$

これらの関係を用いると, トービット分析に対する対数尤度は次のように定義される。ただし積の形式 ($y_i = 0$ および $y_i > 0$ のケースについて, 被説明変数 y_i についての確率をかけあわせる操作を行う) を最初に求めて, この積について対数をとった値について示している。

$$\ln L(\beta, \sigma) = \sum_{i \in (y_i = 0)} \ln \Phi(-x^T\beta/\sigma) + \sum_{i \in (y_i = y_i^*} [\ln \phi[(y_i - x^T\beta)/\sigma] - \ln \sigma] \tag{10}$$

モデルの係数 a_i と分散 σ^2 は, この対数尤度を最大にするように決定される。

2.3 切断分布における回帰:サンプルセレクション分析

サンプルセレクション分析というのは, トービット分析を拡張したものであり, 見かけ上は調査から除外される対象ではあるが, このような除外されたサンプルを含ませることにより, 事象をより分かりやすく把握することができる方法である。例えば, 女性のサンプルについて, 職をもつ場合を $y_1 = 1$, 職についていない場合を $y_1 = 0$ とし, これらの区別が子供の数を示す変数 x により説明されるとする。同時に, 女性の賃金 y_2 とキャリア z との関係を分析する場合も含ませるなどのケースに相当する。賃金が正であるケースは, 当然職をもった女性のサンプルだけである。サンプルセレクション分析は数式により, k 番目のサンプルについて次のように記述することができる。

$$y_{1k}^* = Ax_k^T + \varepsilon_k, A = (a_1, a_2, ..., a_n), x_k = (x_{1k}, x_{2k}, ..., x_{nk}), k = 1, 2, ..., M \tag{11}$$

$$y^*_{2k} = Bz^T_k + \eta_k, A = (b_1, b_2, ..., b_m), z_k = (z_{1k}, z_{2k}, ..., z_{nk}), k = 1, 2, ..., M \tag{12}$$

$$y_{2k} = \begin{cases} y^*_{2k} & \text{if } y_{1k} > 0; \\ 0 & \text{if } y^*_{1k} \leq 0 \end{cases} \tag{13}$$

このモデルにおいて観測可能な変数は y_{1k} の符号と, y^*_{1k} の値が正の場合についてのみ観測される y_{2k} と z_k との対である。ここで,$\varepsilon_k, \eta_k)$ は正規乱数であると仮定しこれらの平均はゼロで, 分散がそれぞれ σ^2_1, σ^2_2 であり, これらの間の共分散が σ_{12} であると仮定する。

$y^*_{1k} \leq 0$ となる確率は, トービット分析と場合と同様に計算される (記述は省略する)。$y^*_{1k} > 0$ において $y_{2k} = y^*_{2k}$ となる対数尤度は, 次のように計算される。

$$Prob(y^*_{1k} > 0 | y_{2k} = y^*_{2k}) = \int p(\varepsilon_{1i}, y_{2k} - Bz^T_k d\varepsilon_k) \tag{14}$$

ここで, $p(.)$ は平均が $(0, 0)$ で共分散行列が $V = [\sigma_{ij}]$ により与えられる二変量正規分布の密度関数である。この対数尤度に式に示した対数尤度を加えたものが, サンプルセレクション分析における対数尤度となる。もとのサンプルセレクション分析において, トービット分析の場合と同様に, この対数尤度関数を最大にするように係数 a_i, b_i および共分散 $\sigma^2_1, \sigma^2_2, \sigma_{12}$ を決めていけばよい。

2.4 KB 手法の適用

これまで述べてきたトービット分析, サンプルセレクション分析において KB 手法を適用する概要は, 以下のようにまとめられる。主な点は通常の回帰分析における説明変数を KB 変数に置き換えることであり, 次のような対応関係になる。

$x_{1k}, x_{2k}, ..., x_{nk} \rightarrow (s_{1k}, s_{2k}, ..., s_{Mk})$

回帰式は次のようになる。

$$y_k = A(s_{1k}, s_{2k}, ..., s_{Mk})^T, A = (a_1, a_2, ..., a_M) \tag{15}$$

これより分かるように, 回帰係数は KB 変数についての係数となるので, その個数はサンプルの総数に等しい。

3 KB 手法におけるサンプル間変数共有

3.1 KB 手法により得られる尤度の性質

以下では KB 手法により得られる非線形関数近似の性質のほかに, サンプルの間で情報が共有されることにより, 尤度が従来手法と比較して, 更に改善されることを説明する。最初に, 簡単な事例によりこの現象を説明する。

話を分かりやすくするために, 人工的に与えたデータに対するトービット分析分析における比較をとりあげる。データ生成には被説明変数である y_k と説明変数 x_{1k}〜x_{nk} を対として与える必要がある。被説明変数は,$y_k > 0$ および $y_k \leq 0$ となる場合を含んでいるが, 自然な仮定として被説明変数の値 y_k が負となるケースは相対的に変数 x_{ik} が小さい場合であり, y_k が正である場合には説明

変数 x_{1k}〜x_{nk} の値が相対的に大きくなるように設定しておく。変数の個数が増加するにしたがって線形回帰と非線形回帰分析の差が大きくなると予想されるので変数の個数を 1 個および 2 個であるケースを仮定する。また, 説明変数の非線形変換が含まれていることが仮定されるので, 以下では次の 6 つのケースを仮定する。Case 1 から Case 3 までは変数が 1 つの場合であり Case 4 以降は変数が 2 つである。

Case 1:$y = x_1^2$
Case 2:$y = \log(2x_1)$
Case 3:$y = \sin(x_1/1.2)$
Case 4:$y = x_1^2 x_2^2$
Case 5:$y = \log(2x_1)\log(2x_2)$
Case 6:$y = \sin(x_1/1.2)\sin(x_2/1.2)$

説明変数の範囲を $x_{ik} = 0$〜2 としておく。これらの仮定に加えて, 被説明変数に, 平均がゼロ分散が $\sigma = 0.1^2$ である正規乱数 ($N(0,\sigma)$ と表記する) を加えておく。サンプルの個数は y_k が負となるケース, 正のケースを, それぞれ 6, 14 としておく。

表 1 には手法ごとのモデル分析の精度の比較分析の結果を示している。この表 1 において対数尤度 L_R, L_K は, それぞれ従来の線形の回帰モデルによる結果, 線形回帰を基礎とした KB 手法による結果を意味している。なおモデルが完全にもとのデータを再現している場合には, 推定誤差は加えられた正規乱数 $N(0,\sigma)$ そのものとなり, これが対数尤度の定義の中身となるので, これを求めて表に L_T として示している。この表より分かるように, 従来の線形のモデルを仮定した場合の分析は一般に良くないが, KB 手法を用いた分析では極めて良好であることが分かる。更に, この KB 手法により得られる対数尤度は, 従来手法により得られる最大尤度の値 L_T を上回っている。これは回帰モデル $y = f(x) + \varepsilon$ を仮定して関数 $f(x)$ を推定することを前提とした場合には, 得られない数値である。したがって KB 手法のもとでは。関数 $f(x)$ だけではなく, その乱数の成分も含めて $f(x) + \varepsilon$ を観測値として近似していることが予測される。

表 1. トービット分析の比較結果

cases	L_T	L_R	L_K	cases	L_T	L_R	L_K
Case 1	10.7	9.6	14.8	Case 4	10.7	-26.5	38.0
Case 2	10.7	2.3	20.0	Case 5	10.7	-12.2	24.0
Case 3	10.7	-17.2	7.3	Case 6	10.7	-11.7	3.6

3.2 サンプル間変数共有による説明

一般にモデル分析におけるデータの共有 (情報共有として整理されている) による残差の減少は, いくつかの分野で観測されている。以下では, 経済分野でよく知られている現象を参考にする [14]-[17]。情報共有は, 企業の間での商品の製造販売について議論されるモデルである。その概要は, 卸売り業者が製造業者に時刻 t における注文 Y_t を出す場合に, 予測した需要 D_t を伝えるだけではなく, 時刻 $t-1$ における予測と実現との差異 ε_{t-1} も伝えることにより, 製造数量推定の平均と分散が ε_{t-1} に比例して改善され, 無駄な製造が回避されるとする議論である [14]-[16]。この議論の対極として存在するモデルでは, 卸売り業者が過大な需要見積もりを繰り返すことにより, 製造数量の変動が拡大される現象も示されている [17]。

本論文では, このような企業の間の情報共有を擬似的にサンプルの間に当てはめる議論であり, 最低限の類似性だけを用いることにする。この場合, 従来の回帰分析と KB 手法による回帰分析におけるサンプルの, データの利用の違いに注目する。まず従来の回帰分析では, 一般に y を説明変数 x の非線形関数 $f(x)$ として関数 $f(.)$ の推定を行うが, その基本は逐次的な近似である。すなわち, あるサンプル k のデータとして被説明変数 y_k と説明変数 x_k の組が与えられた場合に, このサンプルについて $y_k = f(x_k)$ を, できるだけ実現する方向に関数の修正が行われる。しかしながら, 異なるサンプルの間では説明変数の値は独立のものとして用いられる。このような逐次近似を, すべてのサンプルについて繰り返して, 十分な推定誤差が得られるまで繰り返すことになる。もちろん最小 2 乗法では, いちどの計算が完了するが, 手法の基本は同じである。

一方 KB 手法においては, あるサンプル k についての回帰式を成立させる問題を解く場合においては, その変数としては, 通常の説明変数ではなく, KB 変数が用いられる。したがって, サンプル k 以外の関数近似の結果も, このサンプル k についての回帰分析の計算に含まれることになる。例えば y_1 の回帰式に含まれる KB 変数 s_{12} は, y_2 の回帰式に含まれる KB 変数 s_{21} と同じである。したがって, 回帰式の逐次近似を仮定した場合には, この KB 変数の値は規定の数値として被説明変数から差し引かれるので, この分だけ残差が減少すると考えられる。

3.3　尤度を有効利用した変数個数の制限

これまで述べたように, 非線形の回帰分析を効率よく行う方法として KB 手法は有効であることは明らかであるが, 一方では, その変数の個数が推定に用いるサンプルの個数に比例して多くなる問題がある。特に本論文で議論するトービット分析およびサンプルセレクション分析は, 基本的には非線形方程式を解く問題に帰着されるので, KB 変数の増加とともに解を得るまでの時間は増大する

したがって 1 つの考え方として, 回帰分析の精度をやや犠牲にしても, KB 変数の個数を制限する方法論が考えられる。この変更は極めて簡単であり, KB 変数の中からランダムに少数の変数を選択して, この制限された変数のもとで KB 手法にもとづく切断分布への回帰分析を適用することである。しかしながら, このような KB 変数個数の制限が, どのように影響するかを分析しておく必要がある。本論文ではこの変数個数の制限の影響を, 応用例において検討する。

4　応用例

4.1　人工データを用いたシミュレーション

以下では本論文でとりあげる切断分布を含むケースの回帰分析において KB 手法を適用する方法について, まず人工的に生成したデータを用いたシミュレーションにより性能を評価してみる。なお, 本論文ではすでにトービット分析における比較分析の例は 3.1 節において示しているので, 以下ではサンプルセレクション分析についてのみ, 人工的なデータを用いた場合のそれぞれの手法の推定結果の比較分析を行う。まず第 1 段階のサンプル y_{1k}, x_{ik} を選択する部分については, 3.1 節において定義したトービット分析分析のためデータ作成手法を用いることにする。次に第 2 段階の回

帰モデルについては, ここでとりあげているトービット分析の場合と同様に, 定義することにする。すなわち変数の個数を 1 つおよび 2 つに限定してこれらを相互に組みあわせて, 次のようなケースを生成する。なお第 1 段階の y_{1k}, x_{ik} についてのデータ作成はトービット分析と同じであると仮定するので説明は省略する。Case 1 から Case 3 までは変数が 1 つの場合であり Case 4 以降は変数が 2 つである。

Case 1:$y_2 = z_1^2$
Case 2:$y_2 = \log(2z_1)$
Case 3:$y_2 = \sin(z_1/1.2)$
Case 5:$y_2 = z_1^2 z_2^2$
Case 6:$y_2 = \log(2z_1)\log(2z_2)$
Case 7:$y_2 = \sin(z_1/1.2)\sin(z_2/1.2)$

これらの仮定に加えて, 被説明変数に, 平均がゼロ, 分散が $\sigma^2 = 0.1^2$ である正規乱数を加えておく。

表 2 には比較分析の結果を示している。この表 1 における対数尤度 L_K, L_R などの定義はトービット分析の場合と同様である。この表より分かるように, 本論文で導入した KB 手法による分析の結果は, 従来の線形のモデルを仮定した場合の分析結果より, 極めて良好であることが分かる。

表 2. サンプルセレクション分析の結果比較

cases	L_T	L_R	L_K	cases	L_T	L_R	L_K
Case 1	11.2	10.2	12.2	Case 4	11.2	-18.5	28.0
Case 2	11.2	10.4	11.5	Case 5	11.2	-9.9	18.5
Case 3	11.2	5.7	7.7	Case 6	11.2	-5.3	1.9

4.2 German Creit を用いた性能の相互比較

次に, 現実に観測されるデータを用いて, 本論文で述べる KB 手法による切断分布を含む回帰分析の比較を行う。用いるデータはドイツの消費者ローン会社で実施された 1000 名を対象にした貸付審査の結果データであり, 貸付を拒否された 300 名のデータと, 貸付された 700 名のデータからなる [18]。貸付を希望する顧客のデータを入力して, これらの審査結果の違いを出力するような説明ルールを構成することが課題である。データの項目は,7 つの数値データと,13 個のカテゴリカルデータとからなる。表 3 にこれらの項目の概要を示す。このデータセットから回帰分析に適するように,y_k が負となるサンプルを 10 個, 正となるサンプルを 20 個, データセットからランダムに選択している。

表 3. 消費者ローン顧客データの概要

番号	内容	番号	内容	番号	内容
q_0	貸付可否	q_7	有職期間	q_{14}	他の借入
q_1	チェック種類	q_8	貸付利子	z_{15}	家の所有
q_2	貸付期間	q_9	既婚未婚	q_{16}	借入口数
q_3	返済履歴	q_{10}	保証人	q_{17}	職種
q_4	借入目的	q_{11}	居住年数	q_{18}	扶養家族
q_5	貸付金額	q_{12}	資産種類	q_{19}	電話所有
q_6	預金金額	q_{13}	年齢	q_{20}	外国人か

(1) トービット分析分析の比較

トービット分析分析においては, 表 3 に示す変数を次のように回帰モデルの変数などと対応させた分析を行っている。この目的は, 貸付された拒否された顧客の特性分析である。

$q_0 \to y$

$q_1, q_2, q_3, q_5, q_15 \to x_1, x_2, x_3, x_4, x_5$

この分析の結果は, 次のようになる。ここで示す L_K, L_R は, 人工データに対するトービット分析などの場合と同様の定義である。

$L_K = -8.0, L_R = -22.1$

この結果より分かるように, 本論文で導入した KB 手法による分析の結果は, 従来の線形のモデルを仮定した場合の分析結果より, 極めて良好であることが分かる。

(2) サンプルセレクション分析の比較

サンプルセレクション分析においては, 直前に示したトービット分析の第 1 段階に加えて第 2 段階のモデルにおいて, 次のように回帰モデルとの変数の対応関係を仮定する。この目的は, 貸付された顧客の特性分析である。

$q_0 \to y_1, q_5 \to y_2$

$q_1, q_2, q_3, q_5, q_15 \to x_1, x_2, x_3, x_4, x_5$

$q_2, q_8, q_15, q_17, q_19 \to z_1, z_2, z_3, z_4, z_5$

この分析の結果は, 次のようになる。

$L_K = -18.8, L_R = -54.2$ この結果より分かるように, 本論文で導入した KB 手法による分析の結果は, 従来の線形のモデルを仮定した場合の分析結果より, 極めて良好であることが分かる。

4.3 KB 変数個数の削減

次に KB 手法による尤度の改善を用いて KB 変数を削減するため, この効果を用いることについて検証する。この場合, 人工的なデータを用いた場合には, 恣意的な操作が入る可能性があるので, 前節で用いた消費者ローンの事例を用いることにする。検証の方法として単純に, 回帰分析における KB 変数の個数を, ある割合 P にまで削減した場合の, 対数尤度の減少幅 ΔL を求める。なおサンプルセレクション分析の場合には, 説明変数 x_i と z_i との個数を, 同時に割合 P だけ削減することにする。

この結果をまとめたものが表 4 である。この表から分かるように, 変数個数を削減するにしたがって, 対数尤度は悪化しており, 約 6 割を削減 (4 割だけ残す) した場合には, かなり性能が悪化していることが分かり, トービット分析においては, 通常の線形回帰の結果と大差が無いものとなっている。これらの検証結果を参考にして,KB 手法による分析における計算の簡素化を行える可能性がある。

表 4. KB 変数削減による対数尤度の悪化 ΔL

cases	$P = 0.15$	$P = 0.35$	$P = 0.50$	$P = 0.7$
Tobit	0.8	1.0	4.3	7.0
Samp sel	1.8	2.4	5.2	6.0

5 むすび

本論文では, 切断分布をともなう回帰分析における KB 手法の適用し精度や課題を議論した。特に KB[手法においては, サンプルの間の変数共有により尤度が改善されることを示し, これを用いた変数個数の削減の可能性を議論した。応用例として, 人工的に与えたデータの解析と, 実際観測される消費者ローンの問題へと適用し, 本手法の有効性を考察した。

今後の課題として, 多方面における KB 手法の適用可能性の分析などがあり, 検討を進める予定である。

参考文献

[1] B.Scholkopf and A.Smola, “ Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Computation*, vol.10,pp.1299-1319,1998.

[2] H.Kwon and N.M.Nasrabadi, “ Kernel matched subspace detectors for hyperspectral target detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28,no.2,pp.178-194,2006.

[3] G.Baudat and E.Anouar, “Generalized disciminant analysis using a Kernel approach,” *Neural Computing*,vol.12,pp.2385-2404,2000.

[4] J.Zhu and T.Hastie, “ Kernel logistic regression and the import vector machine,” *Journal of Computation and Graphical Statistics*, vol.14,no.1,pp.185-205,2005.

[5] G.Kimeldorf and G.Wahba, “ Some results on Tchebycheffian spline functions,” *Journal of Mathematical Analysis and Applications*, vol.33,pp.82-95,1971.

[6] J.Tobin, “ Estimation of relationship for limited dependent variables,” *Econometrica*, vol.26,pp.24-36,1958.

[7] T.Amemiya, “ Multivariate regression and simultaneous equation models when dependent variables are truncated normal,” *Econometrica*, vol.42,pp.999-1012,1974.

[8] 縄田和満, “ トービット・モデルの金融資産分析への応用について,”大蔵省財政金融研究所ファイナンシャルレビュー, June-1992 号 pp.1-19,1992.

[9] 牧厚志, 古川彰, 渡辺真, 一河信行, 伊藤潔, “ 家計における金融資産選択行動–Tobit Model における資産選択モデルの計測,”郵政研究レビュー,no.1,pp.55-118,1992.

[10] D.Madden, “Sample selection versus two-part models revisited:The case of female smoking and drinking,” HEDG working paper 06/12,pp.1-32, University of York, 2006.

[11] J.J.Heckman, “The common structure of statistical models of truncation: Sample selection and limited dependent variables and a sample estimatitor for such models,” *Annals of Economic and Social Measurement*, vol.5,pp.475-492, 1976.

[12] J.J.Heckman, “Dummy endogenous variables in a simultaneous equation systems,” *Econometrica*, vol.48,pp.931-959, 1978.

[13] 時永祥三, 譚康融, “切断分布をともなう回帰分析における Kernel-based 手法の適用,” 信学技報,NLP2009-175,pp.97-102,2010.

[14] H.Lee,K.C.So and C.S.Tang, “The value of information sharing in a two-level supply chain,” *Management Science*, vol.46,no.5,pp.626-643,2000.

[15] S.Raghunathan, “Information sharing in a supply chain:A note on its value when demand is nonstaritionary,” *Management Science*, vol.47,no.4,pp.605-610,2001.

[16] 岸川 善紀, 時永 祥三, “企業間連係における情報共有のモデル分析とその応用-予測と情報共有コストを中心として”, 経営情報学会論文誌、vol.13,no.1,pp.58-77,2004.

[17] H.L.Lee,V.Padmnabhan and S.Whang, “Information distortion in a supply chain: The bullshiw effect,” *Management Science*, vol.43,no.4,pp.546-558,1997.

[18] *http://www.liacc.up.pt/ML/statlog/datasets/german/german.doc.html*

時永　祥三〔九州大学大学院経済学研究院　教授〕
譚　　康融〔久留米大学経済学部　教授〕