

正準変量のカーネルマッチングによるデータ融合法

Data Fusion Method with Kernel Matching for Estimated Canonical Variables

光廣 正基 *1*2 星野 崇宏 *3*4
Masaki MITSUHIRO Takahiro HOSHINO

*1 (株) 日経リサーチ
Nikkei Research Inc.

*2 慶應義塾大学大学院経済学研究科
Graduate School of Economics, Keio University

*3 慶應義塾大学経済学部
Department of Economics, Keio University

*4 理化学研究所革新知能統合研究センター
RIKEN Center for Advanced Intelligence Project

To clarify the relationship between variables obtained from different sources, we integrate two source data by statistical matching. In this method, missing values of certain data are predicted from covariates and observed values of another data. However, with a large number of covariates, it is difficult to find objects with the same or close covariates. For this reason, statistical matching after reducing covariates in the low-dimensional subspace is useful. In this manuscript, we propose a data fusion method combining kernel canonical correlation analysis and kernel matching. This proposed method can fusion two data sets by kernel matching for estimated canonical variables in the common space reflecting the relationship between covariates and outcome variables.

1. はじめに

回答者や調査モードなどが異なる複数の多変量データを統合する方法として、統計的マッチングやモデルベースの手法がある [Kamakura 97]. 中でも、統計的マッチングは、あるデータの欠測しているアウトカム変数に、そのアウトカム変数が観測されているもう一方のデータから、共変量が同じもしくは一番近い個人の値を割り当てる方法であり、共変量が複数ある場合はマハラノビスマッチング法がよく使用される。

マハラノビスマッチング法は、共変量の数が多い場合は次元の呪いの影響を受けることが知られており [Rubin 78] [Stuart 10], この欠点を解決するため、マッチングする前に共変量を次元縮約することが多い。共変量の縮約方法には、正準相関分析法 [Hotelling 36] が挙げられるが、2つのデータセットの個人が同一である必要があったり [Yu 13], 個人が異なっていたとしても教師や事前情報により個人間の関係がわかっている必要がある [Yan 07] [Shimodaira 14]. さらに、カーネル正準相関分析法 [Akaho 06] を使用すれば、共変量とアウトカム変数の非線形な関係も捉えることが可能である。

また、欠測値にはもう一方のデータで観測された値しか代入できないため、2つのデータの共変量に偏りがある場合に、マハラノビスマッチング法の予測精度が落ちてしまう。そこで、カーネルマッチング法により、各個人との重みを推定し、アウトカム変数の重み付き平均を計算することで、観測していない値も代入可能となる [Fan 93] [Heckman 98].

本研究では、個人の異なる2つの多変量データの統合するため、カーネル正準相関分析法とカーネルマッチング法を組み合わせたデータ融合法を提案する。この手法は、各データの共変量とアウトカム変数の非線形な関係を反映した共通の低次元空間に共変量を縮約し、推定された正準変量に対して、カーネルマッチング法を適用することで、マハラノビスマッチング法の欠点を改善する。

2. 提案手法

ソースの異なる2つのデータ行列の一方を $\mathbf{X}_1 = (x_{ij1})$ ($i = 1, 2, \dots, n_1; j = 1, 2, \dots, p^{\text{cov}} + p^{\text{out}}$) とする。さらに $\mathbf{X}_1$ は共変量 $\mathbf{Z}_1 = (z_{ij1})$ ($i = 1, 2, \dots, n_1; j = 1, 2, \dots, p^{\text{cov}}$) と、観測されているアウトカム変数 $\mathbf{Y}_{11} = (y_{ij11})$ ($i = 1, 2, \dots, n_1; j = 1, 2, \dots, p_1^{\text{out}}$)、欠測しているアウトカム変数 $\mathbf{Y}_{12}^* = (y_{ij12}^*)$ ($i = 1, 2, \dots, n_1; j = 1, 2, \dots, p_2^{\text{out}}$) に分割できる。もう一方のデータ行列も同様に定義でき、2つのデータ行列を統合すると $\mathbf{X} = (x_{ij})$ ($i = 1, 2, \dots, n; j = 1, 2, \dots, p^{\text{cov}} + p^{\text{out}}$) となる。この統合データを作成するために、欠測している $\mathbf{Y}_{12}^*, \mathbf{Y}_{21}^*$ に推定値となる $\hat{\mathbf{Y}}_{12}, \hat{\mathbf{Y}}_{21}$ をそれぞれ代入し、 $\mathbf{X}$ は欠測のない完全データとなる。

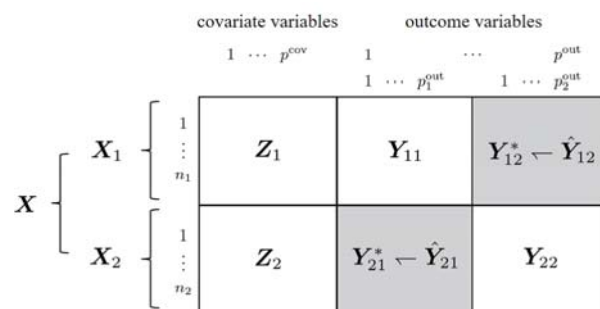


Figure 1: ソースの異なる2つの多変量データ $\mathbf{X}_1, \mathbf{X}_2$

データ行列 $\mathbf{X}_1$ における $\mathbf{Z}_1$ と $\mathbf{Y}_{11}$ のカーネル正準相関分析と、データ行列 $\mathbf{X}_2$ における $\mathbf{Z}_2$ と $\mathbf{Y}_{22}$ のカーネル正準相関分析を同時に行い、共通の低次元空間で推定される共変量側の正準変量に対して、カーネルマッチング法を適用するデータ融合法を提案する。この手法により、個体異なるデータ同士の統合が可能となる。

2.1 STEP1: 次元縮約

非線形関数 $\phi_z(\cdot), \phi_{y_1}(\cdot), \phi_{y_2}(\cdot)$ を用いて, データ行列 $\mathbf{X}_1, \mathbf{X}_2$ の共変量とアウトカム変数を高次元特徴空間に写像し, 特徴ベクトルを抽出する.

$$\begin{aligned} \mathbf{z}_i &\mapsto \phi_z(\mathbf{z}_i), \\ \mathbf{y}_{i11} &\mapsto \phi_{y_1}(\mathbf{y}_{i11}), \\ \mathbf{y}_{i22} &\mapsto \phi_{y_2}(\mathbf{y}_{i22}). \end{aligned}$$

推定したい共変量側の正準変量を $u(\mathbf{z}_i)$, アウトカム変数側の正準変量をそれぞれ $v(\mathbf{y}_{i11}), v(\mathbf{y}_{i22})$ とする. このとき, 中心化した特徴ベクトル $\tilde{\phi}_z(\mathbf{z}_i), \tilde{\phi}_{y_1}(\mathbf{y}_{i11}), \tilde{\phi}_{y_2}(\mathbf{y}_{i22})$, および中心化したカーネル関数 $\tilde{k}_z(\mathbf{z}_i, \cdot), \tilde{k}_{y_1}(\mathbf{y}_{i11}, \cdot), \tilde{k}_{y_2}(\mathbf{y}_{i22}, \cdot)$ を用いて, リプレゼンター定理より, 各正準変量は下記の通りである.

$$\begin{aligned} u(\mathbf{z}_i) &= \tilde{\phi}_z(\mathbf{z}_i)' \mathbf{A}_i \\ &= \tilde{\phi}_z(\mathbf{z}_i)' \sum_{j=1}^{n_k} \tilde{\phi}_z(\mathbf{z}_j) \mathbf{c}_{A_j}' \\ &= \sum_{j=1}^{n_k} \tilde{k}_z(\mathbf{z}_i, \mathbf{z}_j) \mathbf{c}_{A_j}', \\ v(\mathbf{y}_{i11}) &= \tilde{\phi}_{y_1}(\mathbf{y}_{i11})' \mathbf{B}_{i1} \\ &= \tilde{\phi}_{y_1}(\mathbf{y}_{i11})' \sum_{j=1}^{n_1} \tilde{\phi}_{y_1}(\mathbf{y}_{j11}) \mathbf{c}_{B_{1j}}' \\ &= \sum_{j=1}^{n_1} \tilde{k}_{y_1}(\mathbf{y}_{i11}, \mathbf{y}_{j11}) \mathbf{c}_{B_{1j}}', \\ v(\mathbf{y}_{i22}) &= \tilde{\phi}_{y_2}(\mathbf{y}_{i22})' \mathbf{B}_{i2} \\ &= \tilde{\phi}_{y_2}(\mathbf{y}_{i22})' \sum_{j=1}^{n_2} \tilde{\phi}_{y_2}(\mathbf{y}_{j22}) \mathbf{c}_{B_{2j}}' \\ &= \sum_{j=1}^{n_2} \tilde{k}_{y_2}(\mathbf{y}_{i22}, \mathbf{y}_{j22}) \mathbf{c}_{B_{2j}}'. \end{aligned}$$

ここで, $\mathbf{A} = (\mathbf{A}_1', \mathbf{A}_2', \dots, \mathbf{A}_n')'$ は共変量から得られる高次元特徴ベクトルの負荷量を表し, $\mathbf{B}_1 = (\mathbf{B}_{11}', \mathbf{B}_{21}', \dots, \mathbf{B}_{n_11}')'$ や $\mathbf{B}_2 = (\mathbf{B}_{12}', \mathbf{B}_{22}', \dots, \mathbf{B}_{n_22}')'$ はアウトカム変数から得られる高次元特徴ベクトルの負荷量である. さらに, $\mathbf{C}_A = (\mathbf{c}_{A1}', \mathbf{c}_{A2}', \dots, \mathbf{c}_{An}')'$ は共変量側のカーネル関数に対する負荷量であり, $\mathbf{C}_{B_1} = (\mathbf{c}_{B_{11}}', \mathbf{c}_{B_{12}}', \dots, \mathbf{c}_{B_{1n_1}}')'$ と $\mathbf{C}_{B_2} = (\mathbf{c}_{B_{21}}', \mathbf{c}_{B_{22}}', \dots, \mathbf{c}_{B_{2n_2}}')'$ はアウトカム変数側のカーネル関数に対する負荷量である.

共変量側の中心化した特徴ベクトルをデザイン行列 $\tilde{\Phi}_Z = (\tilde{\phi}_z(\mathbf{z}_1)', \tilde{\phi}_z(\mathbf{z}_2)', \dots, \tilde{\phi}_z(\mathbf{z}_n'))'$ で表し, アウトカム変数側も同様とする. このとき, 高次元特徴ベクトルの負荷量は, データ次元の負荷量 $\mathbf{C}_A, \mathbf{C}_{B_1}, \mathbf{C}_{B_2}$ を用いて, 下記のように表す.

$$\begin{aligned} \mathbf{A} &= \tilde{\Phi}_Z \mathbf{C}_A, \\ \mathbf{B}_1 &= \tilde{\Phi}_{Y_1} \mathbf{C}_{B_1} \\ \mathbf{B}_2 &= \tilde{\Phi}_{Y_2} \mathbf{C}_{B_2}. \end{aligned}$$

高次元な負荷量を特徴ベクトルの内積となる中心化グラム行列を $\tilde{\mathbf{K}}_Z = \tilde{\Phi}_Z' \tilde{\Phi}_Z, \tilde{\mathbf{K}}_{Y_k} = \tilde{\Phi}_{Y_k}' \tilde{\Phi}_{Y_k}$ とする. 共変量側の

中心化グラム行列 $\tilde{\mathbf{K}}_Z$ とアウトカム変数側の中心化グラム行列 $\tilde{\mathbf{K}}_Y = \text{diag}(\tilde{\mathbf{K}}_{Y_1}, \tilde{\mathbf{K}}_{Y_2})$ が与えられたとき, データ行列 $\mathbf{X}_1$ でのカーネル正準相関分析とデータ行列 $\mathbf{X}_2$ でのカーネル正準相関分析を同時に行い, 共変量側の正準変量を共通の低次元空間に縮約する. このとき, アウトカム変数側の負荷量を $\mathbf{C}_B = (\mathbf{C}_{B_1}', \mathbf{C}_{B_2}')'$ とすると, 各パラメータを推定する目的関数は下記の通りである.

$$\begin{aligned} f(\mathbf{C}_A, \mathbf{C}_B | \tilde{\mathbf{K}}_Z, \tilde{\mathbf{K}}_Y, \lambda_A, \Lambda_B) \\ = \left\| \tilde{\mathbf{K}}_Z \mathbf{C}_A - \begin{pmatrix} \tilde{\mathbf{K}}_{Y_{11}} \mathbf{C}_{B_1} \\ \tilde{\mathbf{K}}_{Y_{22}} \mathbf{C}_{B_2} \end{pmatrix} \right\|^2 \\ + \lambda_A \mathbf{C}_A' \tilde{\mathbf{K}}_Z \mathbf{C}_A + \Lambda_B \begin{pmatrix} \mathbf{C}_{B_1}' \tilde{\mathbf{K}}_{Y_1} \mathbf{C}_{B_1} \\ \mathbf{C}_{B_2}' \tilde{\mathbf{K}}_{Y_2} \mathbf{C}_{B_2} \end{pmatrix}, \end{aligned} \quad (1)$$

subject to

$$\begin{aligned} \mathbf{C}_A' (\tilde{\mathbf{K}}_Z^2 + \lambda_A \tilde{\mathbf{K}}_Z) \mathbf{C}_A &= \mathbf{I}_d, \\ \mathbf{C}_B' (\tilde{\mathbf{K}}_Y^2 + \Lambda_B \tilde{\mathbf{K}}_Y) \mathbf{C}_B &= \mathbf{I}_d. \end{aligned}$$

ここで, $\lambda_A, \lambda_{B_1}, \lambda_{B_2}$ は正則化パラメータであり, アウトカム変数側の正則化パラメータは $\Lambda_B = \text{diag}(\lambda_{B_1} \mathbf{I}_{n_1}, \lambda_{B_2} \mathbf{I}_{n_2})$ とする. また, $\mathbf{I}_d$ は d 次元の単位行列であり, 正準変量に直交制約が課せられる.

上記の目的関数を制約条件の下で最小化するため, 特異値分解を行う.

$$(\tilde{\mathbf{K}}_Z^2 + \lambda_A \tilde{\mathbf{K}}_Z)^{-\frac{1}{2}} \tilde{\mathbf{K}}_Z \tilde{\mathbf{K}}_Y (\tilde{\mathbf{K}}_Y^2 + \Lambda_B \tilde{\mathbf{K}}_Y)^{-\frac{1}{2}} = \mathbf{S} \mathbf{\Xi} \mathbf{R}',$$

ここで, $\mathbf{S}' \mathbf{S} = \mathbf{R}' \mathbf{R} = \mathbf{I}_n$ は特異ベクトル, $\mathbf{\Xi} = \text{diag}(\xi_1, \xi_2, \dots, \xi_n)$ は特異値であり, ξ_i ($i = 1, 2, \dots, n$) は降順に並べられている. 特異値分解の結果を用いて, パラメータの推定式は下記の通りである.

$$\begin{aligned} \hat{\mathbf{C}}_A &= (\tilde{\mathbf{K}}_Z^2 + \lambda_A \tilde{\mathbf{K}}_Z)^{-\frac{1}{2}} \mathbf{S}_d, \\ \hat{\mathbf{C}}_B &= (\tilde{\mathbf{K}}_Y^2 + \Lambda_B \tilde{\mathbf{K}}_Y)^{-\frac{1}{2}} \mathbf{R}_d. \end{aligned} \quad (2)$$

ここで, $\hat{\mathbf{C}}_A$ と $\hat{\mathbf{C}}_B$ は各グラム行列の d 次元の負荷量の推定値を表す.

2.2 STEP2: マッチング

推定した $\hat{\mathbf{C}}_A$ から, データ行列 $\mathbf{X}_1$ の共変量側の正準変量 $\tilde{\mathbf{K}}_{Z_1} \hat{\mathbf{C}}_A$ とデータ行列 $\mathbf{X}_2$ の共変量側の正準変量 $\tilde{\mathbf{K}}_{Z_2} \hat{\mathbf{C}}_A$ を求め, これらに統計的マッチングを適用し, データ行列 $\mathbf{X}_1, \mathbf{X}_2$ の欠測を補完する.

実際には, $\tilde{\mathbf{K}}_{Z_1} \hat{\mathbf{C}}_A$ をレシピエントデータの共変量, $\tilde{\mathbf{K}}_{Z_2} \hat{\mathbf{C}}_A$ をドナーデータの共変量として, レシピエントデータのアウトカム変数 $\hat{\mathbf{Y}}_{12}$ を推定する. 一方, レシピエントとドナーを入れ替え, $\tilde{\mathbf{K}}_{Z_2} \hat{\mathbf{C}}_A$ をレシピエントデータの共変量, $\tilde{\mathbf{K}}_{Z_1} \hat{\mathbf{C}}_A$ をドナーデータの共変量として, レシピエントデータのアウトカム変数 $\hat{\mathbf{Y}}_{21}$ を推定する.

$\hat{\mathbf{C}}_A$ と $\hat{\mathbf{C}}_B$ において, 次元数が $d > 1$ であれば, マハラノビスマッチング法などが適用でき, $d = 1$ であれば, 最近傍マッチング法やカーネルマッチング法などが適用できる. 本研究では, カーネルマッチング法を使用する.

3. シミュレーション

提案手法と従来手法の欠測しているアウトカム変数の予測精度の比較を行う。提案手法は、推定する正準変量を1次元に設定し、カーネル正準相関分析法によって得られた共変量側の正準変量に対して、カーネルマッチング法を適用する。従来手法には多変量データのマッチング手法で代表的なマハラノビスマッチング法を用いる。

3.1 事前設定

STEP1のカーネル正準相関分析法において、カーネル関数はガウスカーネルを用い、バンド幅は入力値の2乗誤差の中央値とする。正則化パラメータは5-foldクロスバリデーションによって決定する。正則化パラメータの候補は(0.1, 0.5, 1.0)とし、この中から最適な値の組を選ぶとする。

データ行列を g 個に分割したとき、その一つをテストデータとし、残りの $(g-1)$ 個をトレーニングデータとする。 $\tilde{\mathbf{K}}_{\mathbf{Z}_1}^{(g)}$ や $\tilde{\mathbf{K}}_{\mathbf{Y}_1}^{(g)}$ はテストデータの中心化グラム行列を表し、 $\hat{\mathbf{C}}_{\mathbf{A}}^{(-g)}$ や $\hat{\mathbf{C}}_{\mathbf{B}}^{(-g)}$ はトレーニングデータの推定値を表す。クロスバリデーションの評価関数は正準変量間の誤差を用い [Hardoon 09], データ行列 $\mathbf{X}_1$ の正準変量間の誤差とデータ行列 $\mathbf{X}_2$ の正準変量間の誤差の平方和の5回平均とする。

$$\varepsilon(\lambda_{\mathbf{A}}, \lambda_{\mathbf{B}}) = \frac{1}{5} \sum_{g=1}^5 \left(\left\| \tilde{\mathbf{K}}_{\mathbf{Z}_1}^{(g)} \hat{\mathbf{C}}_{\mathbf{A}}^{(-g)} - \tilde{\mathbf{K}}_{\mathbf{Y}_1}^{(g)} \hat{\mathbf{C}}_{\mathbf{A}}^{(-g)} \right\|^2 + \left\| \tilde{\mathbf{K}}_{\mathbf{Z}_2}^{(g)} \hat{\mathbf{C}}_{\mathbf{B}}^{(-g)} - \tilde{\mathbf{K}}_{\mathbf{Y}_2}^{(g)} \hat{\mathbf{C}}_{\mathbf{B}}^{(-g)} \right\|^2 \right).$$

また、STEP2のマッチング法の設定として、カーネルマッチング法ではガウスカーネルを用いてアウトカム変数を予測する。共変量が同じもしくは近い個人を決定するマハラノビスマッチング法では、1度マッチングされたドナーデータの個人は何回も利用できるとする。

3.2 シミュレーションデータ

50個のシミュレーションデータ $\mathbf{X}$ を作成し、2つのデータ行列に分割した後、各データ行列で欠測変数を作成する。共変量はアウトカム変数と非線形な関係を持つような2変数と10変数の $N(0, \mathbf{I}^2)$ に従う正規ノイズから作られる。2つの共変量の生成は下記の通りである。

$$\begin{aligned} z_{i1} &= \theta + \varepsilon_{i1}^{(z)}, \\ z_{i2} &= 5 + 2\theta + 0.8\theta^2 - 3\theta^3 + \varepsilon_{i2}^{(z)}, \\ \theta &\stackrel{\text{i.i.d.}}{\sim} U(-1, 1), \\ \varepsilon_{ij}^{(z)} &\stackrel{\text{i.i.d.}}{\sim} N(0, 0.1^2) \\ (i &= 1, 2, \dots, n; j = 1, 2). \end{aligned}$$

続いて、共変量の θ を用いて、アウトカム変数を4変数作成する。アウトカム変数の生成は下記の通りである。

$$\begin{aligned} y_{i1} &= 1 + \theta + 5\theta^2 + \varepsilon_{i1}^{(y)}, \\ y_{i2} &= 1 - 3\theta + 3\theta^2 + \varepsilon_{i2}^{(y)}, \\ y_{i3} &= 2 + \theta + \theta^4 + \varepsilon_{i3}^{(y)}, \\ y_{i4} &= \theta^2 - 0.5\theta^3 + \varepsilon_{i4}^{(y)}, \\ \varepsilon_{ij}^{(y)} &\stackrel{\text{i.i.d.}}{\sim} N(0, 0.1^2) \\ (i &= 1, 2, \dots, n; j = 1, 2, 3, 4). \end{aligned}$$

生成した 1000×16 のデータ行列 $\mathbf{X}$ を2分割する。分割方法は、ロジスティック関数で得られる確率 $p_i = 1/(1+e^{-\theta})$ に従って、 $n_1 = 500$ の $\mathbf{X}_1$ 、残りを $\mathbf{X}_2$ とし、 $n_2 = n - n_1 = 500$ となる。 $\mathbf{X}_1$ については、 y_{i3}, y_{i4} を欠測させ、 $\mathbf{X}_2$ については、 y_{i1}, y_{i2} を欠測させる。この4つのアウトカム変数の欠測の値を予測する。

3.3 比較結果

50個のシミュレーションデータに対して、提案手法と従来手法を適用し比較する。まず、提案手法とマハラノビスマッチング法の相関係数を比較する。提案手法の真値と推定値との相関係数とマハラノビスマッチング法の真値と推定値の相関係数を計算し、差分をとる。

$$\rho_j = \text{corr}(\mathbf{y}_j, \hat{\mathbf{y}}_j) - \text{corr}_{\text{MM}}(\mathbf{y}_j, \hat{\mathbf{y}}_j).$$

各変数の差分が正の場合($\rho_j > 0$)は提案手法の予測精度が良く、負の場合($\rho_j < 0$)は提案手法の予測精度が従来手法よりも悪いと評価する。Figure 2は相関係数の比較結果である。Figure 2から、提案手法は4つのアウトカム変数すべてにおいて、相関係数の差分が正に大きく、従来手法のマハラノビスマッチング法よりも予測精度が改善されている。

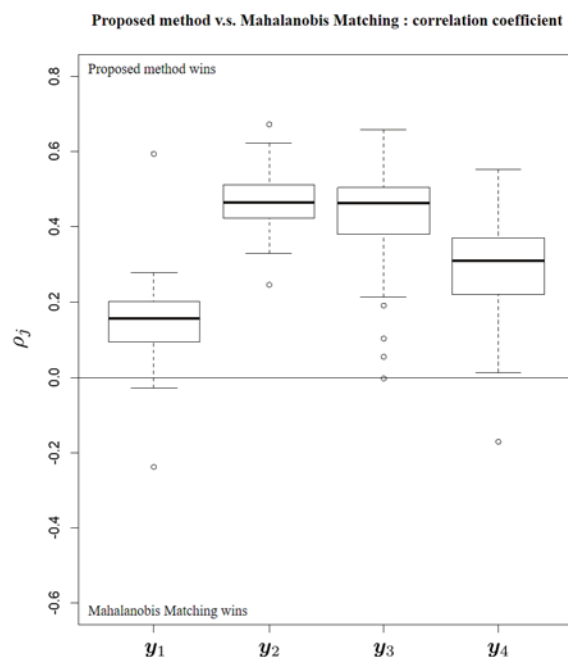


Figure 2: 提案手法とマハラノビスマッチング法の相関係数の差分 ρ_j

続いて、提案手法とマハラノビスマッチング法の平均二乗誤差(MSE)を比較する。各アウトカム変数に対して、MSEを計算し、マハラノビスマッチング法との相対比を計算する。MSEの比が1より小さい場合は提案手法の予測精度が良く、1より大きい場合は提案手法の予測精度が従来手法よりも悪いと評価する。Figure 3はMSEの比較結果である。Figure 3から、提案手法は4つのアウトカム変数すべてが1より小さく、従来手法のマハラノビスマッチング法よりも予測精度が改善されている。

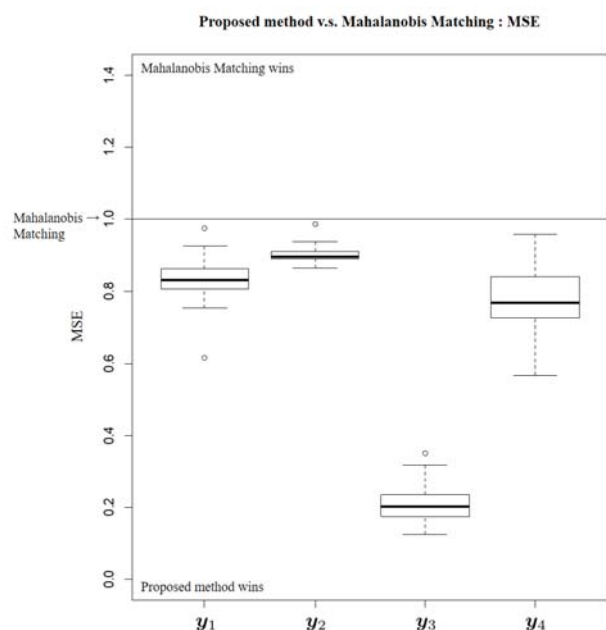


Figure 3: 提案手法とマハラノビスマッチング法の MSE の相対比

4. おわりに

カーネル正準相関分析法とカーネルマッチング法を組み合わせたデータ融合法を提案した。2つのデータ行列のカーネル正準相関分析を同時に行い、共変量側の正準変量については、共通の低次元空間に縮約した。推定された正準変量に対して、カーネルマッチング法を行い、ドナーデータにない値もレシピエントデータのアウトカム変数に代入できるようにした。シミュレーションデータを用いて、従来手法のマハラノビスマッチング法と比較し、提案手法のほうが相関係数が大きく、MSEが小さいという結果となり、予測精度の改善がみられた。

参考文献

- [Akaho 06] Akaho, S. (2006). A kernel method for canonical correlation analysis. arXiv preprint cs/0609071.
- [Fan 93] Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 196–216.
- [Hardoon 09] Hardoon, D. R. and Shawe-Taylor, J. (2009). Convergence analysis of kernel canonical correlation analysis: theory and practice. *Machine learning*, **74**, 23–38.
- [Heckman 98] Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *The review of economic studies*, **65**, 261–294.
- [Hotelling 36] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, **28**, 321–377.
- [Kamakura 97] Kamakura, W. A. and Wedel, M. (1997). Statistical data fusion for cross-tabulation. *Journal of Marketing Research*, 485–498.
- [Rubin 78] Rubin, D. B. (1978). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *ETS Research Report Series*, **1978**.
- [Shimodaira 14] Shimodaira, H. (2014). A simple coding for cross-domain matching with dimension reduction via spectral graph embedding. arXiv:1412.8380.
- [Stuart 10] Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, **25**, 1.
- [Yan 07] Yan, S., Xu, D., Zhang, B., Zhang, H. J., Yang, Q., and Lin, S. (2007). Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, **29**, 40–51.
- [Yu 13] Yu, S., Tranchevent, L. C., De Moor, B., and Moreau, Y. (2013). *Kernel-based data fusion for machine learning*. Springer-Verlag Berlin An.