データベクトル間マッチングの多変量解析とそのクロスバリデーション

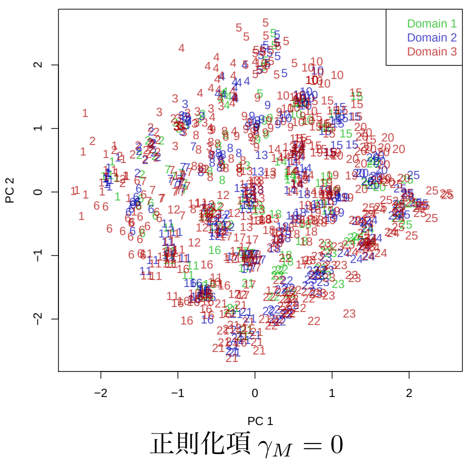
大阪大学　下平英寿

N 個の P 次元データベクトル $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^P$ が与えられていて，データ行列を $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T \in \mathbb{R}^{N \times P}$ とする．二つのデータベクトル $\boldsymbol{x}_i$，$\boldsymbol{x}_j$ 間の関連の強さを事前情報をもとにマッチングウエイト $w_{ij} = w_{ji} \geq 0$ で指定して，対称行列 $\boldsymbol{W} = (w_{ij}) \in \mathbb{R}^{N \times N}$ を定義する．次元削減のため $\mathbb{R}^P$ から $\mathbb{R}^K$ $(K \leq P)$ への線形変換 $\boldsymbol{y}_i = \boldsymbol{A}^T \boldsymbol{x}_i$, $i = 1, \ldots, N$ を行うが，互いに関連の強いベクトルは近くに配置されるような線形変換を求めたい．そこで $\boldsymbol{X}$ と $\boldsymbol{W}$ を観測したとき，マッチング誤差

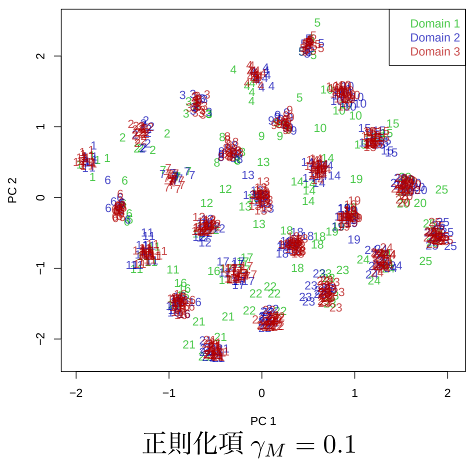
$$\phi = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \| \boldsymbol{y}_i - \boldsymbol{y}_j \|^2$$

を最小にするように行列 $\boldsymbol{A} \in \mathbb{R}^{P \times K}$ を定める．ただし，変換したデータベクトルが縮退しないように制約条件を入れる．Yan et al. (2007) では $\boldsymbol{X}$ から $\boldsymbol{W}$ を定めて同様の最適化問題を考えているが，ここでは一般に他の情報から $\boldsymbol{W}$ を定めている．我々は，この定式化をマッチング相関分析 (Matching Correlation Analysis, MCA) と呼ぶことにする．実はこの定式化が主成分分析 (PCA) や正準相関分析 (CCA) を特殊な場合として含む一般的な多変量解析を与える．Shimodaira (2014) では複数ドメインの場合でもごく単純なコーディングによって MCA に帰着できることを示している．つまり D 個のドメインを考え，各ドメイン $d = 1, \ldots, D$ において n_d 個の p_d 次元データベクトルが与えられるクロスドメイン・マッチング相関分析 (Cross-Domain Matching Correlation Analysis, CDMCA) でも，$N = \sum_d n_d$，$P = \sum_d p_d$ とした MCA へ容易に帰着できる．

本講演では Shimodaira (2015) の結果を紹介する．MCA や CDMCA に正則化項を加えた場合に，その正則化項の大きさをどのように決めるかが問題になる．母集団に相当する未知のマッチングウェイト行列 $\bar{\boldsymbol{W}}$ からのサンプリングによって $\boldsymbol{W}$ が得られたと想定して，$\bar{\boldsymbol{W}}$ に関するマッチング誤差 ϕ^{true} を小さくするような正則化項を求めたい．データベクトルをリサンプリングするかわりにマッチングウェイトをリサンプリングするクロスバリデーションを行って ϕ^{cv} を計算する．$\boldsymbol{W}$ がスパース行列であることを仮定するとき，N と P がある条件をみたして大きくなる漸近理論によって，ϕ^{cv} が ϕ^{true} の不偏推定量になることを示す．



正則化項 $\gamma_M = 0$



正則化項 $\gamma_M = 0.1$

S. Yan, et al. Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29, 40–51, (2007).

H. Shimodaira. A simple coding for cross-domain matching with dimension reduction via spectral graph embedding. arXiv:1412.8380 (2014).

H. Shimodaira. Cross-validation of matching correlation analysis by resampling matching weights. arXiv:1503.08471 (2015).