ガウス過程潜在変数モデルを用いたデータ融合法の提案

（株）日経リサーチ・慶應義塾大学大学院経済学研究科　　光廣 正基
慶應義塾大学経済学部・理化学研究所 AIP センター　　星野 崇宏

マーケティングや経済学など様々な分野において，取得方法が異なることで調査モードや回答者の違いがある複数の多変量データがある場合，同時に観測されない変数間の関係を調べるためには，これらのマルチソースデータを 1 つに統合する必要がある．統計的データ融合 (Kamakura and Wedel, 1997) は，複数の多変量データをシングルソースデータとして統合する手法であり，統計的マッチングや回帰モデル，潜在変数モデルなどがよく使用される．中でも，共変量とアウトカム変数の背後に共通因子や潜在クラスを仮定し，欠測値を予測する手法が提案されているが (Kamakura and Wedel, 2000)，この手法は潜在変数から観測値への写像が線形となっている．

また，観測値と潜在変数との非線形な関係を捉える教師なし学習の手法として，ガウス過程潜在変数モデルが提案されている (Lawrence, 2004, 2005). この手法は主成分分析をガウス過程に拡張したモデルであり，機械学習の分野で応用されている．量的変数をもつ多変量データ y_{ij} $(i = 1, 2, \cdots, n; j = 1, 2, \cdots, p)$ があるとき，潜在変数 $\boldsymbol{z}_i$ によるガウス過程 $f_{(j)}$ から生成される潜在変数モデルは下記となる．

$$f_{(j)} \sim \mathcal{GP}(\mu(\boldsymbol{z}), k(\boldsymbol{z}, \boldsymbol{z}^{'}))$$
$$y_{ij} = f_{(j)}(\boldsymbol{z}_i) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

ここで，$\mu(\boldsymbol{z})$ は潜在変数の平均，$k(\boldsymbol{z}, \boldsymbol{z}^{'})$ はカーネル関数である．もし観測データ y_{ij} が離散値のとき，離散選択モデルに対してもガウス過程を用いることができ，多変量データが質的変数でもガウス過程による潜在変数が推定可能である．

本研究では，個人の異なる 2 つの多変量データをシングルソースデータとして統合するため，教師なし学習のひとつであるガウス過程潜在変数モデルを用い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間の非線形な関係を捉えることができ，推定された潜在変数から欠測値を予測する．両方のデータセットで観測される共変量 $\boldsymbol{x}_i$ $(i = 1, 2, \cdots, n)$ と片方のデータセットでしか観測できないアウトカム変数 $\boldsymbol{y}_{ik}$ $(i = 1, 2, \cdots, n_k; k = 1, 2)$ で構成される 2 つの多変量データが与えられたとき，欠測データを含むガウス過程潜在変数モデルの尤度は下記の通りである．

$$\prod_{i=1}^{n} \int_{\boldsymbol{f}_i} \Big\{ p(\boldsymbol{y}_{i1}|\boldsymbol{f}_i) p(\boldsymbol{x}_i|\boldsymbol{f}_i) \Big\}^{r_i} \Big\{ p(\boldsymbol{y}_{i2}|\boldsymbol{f}_i) p(\boldsymbol{x}_i|\boldsymbol{f}_i) \Big\}^{1-r_i} p(\boldsymbol{f}_i) d\boldsymbol{f}_i$$

ここで，$\boldsymbol{f}_i$ は潜在変数のガウス過程，$r_i \in \{0, 1\}$ は欠測の割り当てを表すインディケータであり，2 つの多変量データを区別している．MCMC を用いてパラメータを推定し，得られた同時分布から欠測部分の値を推定する．

参考文献

[1] Kamakura, W. A. and Wedel, M. (1997). Statistical data fusion for cross-tabulation. *Journal of Marketing Research*, 485–498.

[2] Kamakura, W. A. and Wedel, M. (2000). Factor analysis and missing data. *Journal of Marketing Research*, **37**, 490–498.

[3] Lawrence, N. D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. *Advances in Neural Information Processing Systems*, **16**, 329–336.

[4] Lawrence, N. D. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, **6**, 1783–1816.