臨床試験でヒストリカル対照群を使用するための研究デザイン

西本 博之 [1]
[1] 大阪産業大学 デザイン工学部 情報システム学科
e-mail : nishimoto@ise.osaka-sandai.ac.jp

1　緒言

Historical Control Trial (HCT) とは，進行中の研究における患者の標準治療に対する潜在的な反応を推定するための非ランダム化比較試験のことである [1]．HCT の治療群は前向き研究で、対照群は Real World Data を用いた後ろ向き研究のため，全てのエントリーを新しい治療法に割り付けられる期待があるが，同時対照群ではないため偏りが生じる課題もある [2]．HCT における潜在的バイアスを軽減するために，傾向スコア法を動的割付として用いる試験デザインについて議論する．

2　誤差分布のモデル

HCT の試験デザインに起因するバイアスを明確にするためのモデルとして，誤差分布のモデルを設定した．各母集団の誤差分布は正規分布であり，平均値をゼロと仮定した．ここで，サンプルサイズ，分散，標準誤差の 3 つのパラメータで，誤差分布を $N\big(n, \sigma^2, \frac{\sigma}{\sqrt{n}}\big)$ で表すと，正規分布の加法定理により次式が与えられる．

$$N\left(n+n, \sigma^2+\sigma^2, \sqrt{\frac{\sigma^2+\sigma^2}{n+n}}\right) = N\left(2n, 2\sigma^2, \sqrt{\frac{\sigma^2}{n}}\right). \tag{1}$$

一般化して，母集団のサンプルサイズ $n_0 = \beta \cdot n \quad (\beta \geq 1)$ と定義すると，次式が与えられる．

$$N\left(n_0, {\sigma_0}^2, \sqrt{\frac{{\sigma_0}^2}{n_0}}\right) = N\left(\beta n, \beta\sigma^2, \sqrt{\frac{\beta\sigma^2}{\beta n}}\right) = N\left(\beta n, \beta\sigma^2, \sqrt{\frac{\sigma^2}{n}}\right). \tag{2}$$

式 (1)・式 (2) より，同じ母集団から派生したサブセットでは標準誤差が等しいことが分かる．

一方 HCT のバイアスについては，ヒストリカル対照群を生成することが，選択バイアスを生成することに繋がり，その選択バイアスが治療群と対照群の平均値の差に応じてプラスにもマイナスにも変化するため，共変量の調整が難しいことが分かる．つまりこの選択バイアスを排除するには，ヒストリカル対照群に対して無作為化などの手順が必要であることが分かる．

3　傾向スコアマッチングを用いた動的無作為化

HCT の選択バイアスを除去するテクニックとして，傾向スコアマッチングを用いた動的割付が考えられる．前向きの治療群に入る症例と同じ傾向スコアを持つ過去の症例を動的に割り付ける手順である [2]．具体的には，前向きの治療群に入る症例の傾向スコアに最も近い過去の症例を１つずつ選択する．この動的無作為化により，HCT に起因する選択バイアスだけでなく，様々なバイアスを排除することが期待される．但し，傾向スコアはロジスティック回帰を用いて，複数の要因をマッチングし易いように１つの入院リスクにまとめた値なので，入院リスクが明らかに異なる集団の比較には適さない．例えば，プラセボ群との比較試験は傾向スコアマッチングに適さず，新旧の術式の比較のように傾向スコアが類似したグループの比較試験の方が適している．

4 Historical Control Trials (HCTs)

傾向スコアマッチングを用いた動的無作為化は，HCT に起因する様々なバイアスを除去することができる．しかしながら，無条件に動的無作為化を導入して良いという訳ではない．極端な例としては，ヒストリカルデータと対照群のサンプルサイズが等しい場合，最終マッチング症例で，ヒストリカルデータの残数は 1 例となり，マッチングの選択の自由はなく，マッチング誤差は最大となる．このようにヒストリカル対照群の目標例数と，ヒストリカルデータの症例数との間の条件の明確化が必要である．言い換えると，同じ母集団から派生したサブセットの標準誤差が安定するために必要なサンプルサイズの条件を求めることである．ヒストリカル対照群はヒストリカルデータのサブセットなので，標準誤差は一致する．従ってヒストリカルデータのサンプルサイズが十分に大きければ，マッチング誤差は標準誤差で最小となる．逆にヒストリカルデータのサンプルサイズが小さいと，マッチング時の際の選択の自由は無くなり，マッチング誤差は最大化する．これらの条件から傾向スコアマッチングを用いた動的無作為化では，前向き治療群の目標例数を n_1、ヒストリカルデータのサンプルサイズを n_0 とすると．次式を満足する必要がある．

$$n_0 \geq 2n_1 - 1 \qquad (n_0 \geq n_1)\,. \tag{3}$$

5 結言

ランダム化されていないヒストリカルコントロールグループは比較試験の構成要素として認められない．同時対照群ではないため，選択バイアスを排除できないからである [2]．しかしながら，傾向スコアマッチングを用いた動的無作為化により，ヒストリカルコントロール試験が可能となる．具体的には，前向きの治療群を割り付ける時に，同じ傾向スコアを持つ過去の症例を 1 例ずつ割り付ける．しかしながら，傾向スコアマッチングを用いた動的無作為化では，前向き治療群の目標例数の少なくとも約 2 倍のヒストリカルデータの症例が必要となる．それでもこのテクニックにより，シングルアームの制約から開放され，これまで蓄積されたヒストリカルデータの半分は有効活用される．これはサンプル数の大小に関わらず，同じ母集団から派生したサブセットの標準誤差は一致するという理論に基づいている．マッチング誤差を２種類に分けると，１つは観測データがデータ分布のどの辺りに出現するかという標準偏差で表される誤差がある．前向き治療群に起因するマッチング誤差は，この標準偏差で表される誤差である．２つ目は，その観測データがどれだけ外れてマッチングされるかという標準誤差で表される誤差である．ヒストリカル対照群に起因するマッチング誤差は，この標準誤差で表される誤差である．従って，本稿では同じ母集団から派生したサブセットの標準誤差にフォーカスしたモデルを検討した．しかしながら，このモデルを基本として，その他の因子についても慎重な議論が必要と考える．

参考文献

[1] Ghadessi et al, A roadmap to using historical controls in clinical trials - by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG) Orphanet Journal of Rare Diseases **15**: 69 (2020). 1-19

[2] 厚生労働省医薬局審査管理課長, 医薬審発第 136 号（別添）臨床試験における対照群の選択とそれに関連する諸問題，平成 13 年 2 月 27 日.

Bregman ダイバージェンスで捉える一般化線形モデルの幾何

熊谷 敦也 [1]
[1] 日本大学商学部
e-mail : kumagai.atsuya@nihon-u.ac.jp

1　はじめに

一般化線形モデルを概観し，そこで主要な役割を持つ指数型分布族の性質について Bregman ダイバージェンスの観点でまとめる．

1.1　一般化線形モデル

一般化線形モデル (GLM)[1] は，従属変数が正規分布に従うとした線形モデルの一般化であり，正規分布を指数型分布族へ一般化したものに相当する．確率密度関数 $p(y;\theta) = \exp\left[y\theta - \psi(\theta)\right] p_o(y)$ で表される指数型分布族を考えることにすると，θ は自然母数と呼ばれる未知母数であり，その関数 $\psi(\theta)$ はキュムラント母関数と呼ばれる．指数型分布族には代表的な確率分布の多くが含まれ，キュムラント母関数 $\psi(\theta)$ の形を与えることで確率分布が決まる．$\theta = \mu$, $\psi(\theta) = \sigma^2\theta^2/2$ とすれば正規分布 $N(\mu, \sigma^2)$ に帰着する．一方で $\theta = \lambda$, $\psi(\theta) = \exp(\theta)$ とすればポアソン分布 $Po(\lambda)$ に帰着する．

以下では簡単のため単回帰の場合を考え，独立変数のデータ $\boldsymbol{x} = (x_1 \cdots x_n)'$ と従属変数のデータ $\boldsymbol{y} = (y_1 \cdots y_n)'$ を考える．GLM では，線形予測子 $\alpha + \beta x_i$ が y_i の期待値 μ_i の関数 $g(\mu_i)$ に等しいとする．この関数 $g(\mu_i)$ はリンク関数と呼ばれる．

1.2　指数型分布族と Bregman ダイバージェンス

ベクトル $\boldsymbol{y}$ に関する凸関数 $\phi(\boldsymbol{y})$ があったとする．このとき， 関数 ϕ に基づき $\boldsymbol{y}$, $\boldsymbol{z}$ 間の隔たりの度合いを表す Bregman ダイバージェンス $d_\phi(\boldsymbol{y}, \boldsymbol{z})$ が以下のように導入される：

$$d_\phi(\boldsymbol{y}, \boldsymbol{z}) = \phi(\boldsymbol{y}) - \phi(\boldsymbol{z}) - (\boldsymbol{y} - \boldsymbol{z}) \cdot \nabla\phi(\boldsymbol{z}). \tag{1}$$

確率ベクトル $\boldsymbol{y}$ が従う指数型分布族を特徴づけるキュムラント母関数 $\psi(\boldsymbol{\theta})$ に対して，ルジャンドル変換 $\phi(\boldsymbol{\mu}) = \boldsymbol{\theta} \cdot \boldsymbol{\mu} - \psi(\boldsymbol{\theta})$, $\boldsymbol{\mu} = \nabla\psi(\boldsymbol{\theta})$ および双対な変換を考えたとき，$\boldsymbol{\mu}$ は $\boldsymbol{y}$ の期待値であり，$\boldsymbol{y}$ の $\boldsymbol{\mu}$ からの隔たりは，$\boldsymbol{\theta} = \nabla\phi(\boldsymbol{\mu})$ を用いて $d_\phi(\boldsymbol{y}, \boldsymbol{\mu}) = \phi(\boldsymbol{y}) + \psi(\boldsymbol{\theta}) - \boldsymbol{y} \cdot \boldsymbol{\theta}$ と書かれる．$\psi(\boldsymbol{\theta})$ が特定されれば $d_\phi(\boldsymbol{y}, \boldsymbol{\mu})$ が一意に特定され，その逆も成立する [2]．指数型分布族では，負の対数尤度における母数への依存性を Bregman ダイバージェンスの形で取り出せるという性質がある [2].

2　Bregman ダイバージェンスで捉える一般化線形モデル

通常 GLM では最尤法によって回帰母数の推定を行うが，指数型分布族を考える限り，尤度の最大化は Bregman ダイバージェンスの最小化と等価である [2]．以下ではこの観点で GLM を捉える．

リンク関数が自然母数と等しく $\theta_i = g(\mu_i) = \alpha + \beta x_i$ であるとする．このときリンク関数は正準リンク関数と呼ばれる．$\boldsymbol{y}$ の期待値を $\boldsymbol{\mu}$ とすると，ダイバージェンスを最小化する α, β を推定するための連立方程式は以下のように書ける：

$$\frac{\partial}{\partial\alpha} d_\phi(\boldsymbol{y}, \boldsymbol{\mu}) = (\boldsymbol{\mu} - \boldsymbol{y})'\mathbf{1}_n = 0, \tag{2}$$

$$\frac{\partial}{\partial\beta}d_\phi(\boldsymbol{y},\boldsymbol{\mu})=(\boldsymbol{\mu}-\boldsymbol{y})'\boldsymbol{x}=0. \tag{3}$$

ここで $\mathbf{1}_n$ は n 次元ベクトル $(1,\cdots,1)'$ である．(2)(3) から決まる α, β による $\boldsymbol{\mu}$ の推定値を $\hat{\boldsymbol{y}}$ と書くとその要素は $\hat{y}_i=g^{-1}(\alpha+\beta x_i)$ のように独立変数 x_i に関して非線形となりうるが，この点を除いては線形モデルにおける正規方程式と同様の方程式が得られることが分かる．

仮に $\beta=0$ に固定した場合，これは $\theta_i=\alpha$ というパラメタ数が 1 個だけのモデルを考えることに相当し，これは null モデルと呼ばれることがある．このとき μ_i の推定値は i によらず $\hat{y}_i=\bar{y}=g^{-1}(\alpha)$ （ $\bar{y}=\sum_{i=1}^n y_i/n$ ）から $\alpha=g(\bar{y})$ が求まる．

3　一般化線形モデルの情報幾何的描像

方程式 (2)(3) は， $\mathbf{1}_n$ とデータ $\boldsymbol{x}$ が張る e-平坦な空間に，データ $\boldsymbol{y}$ と 推定値 $\hat{\boldsymbol{y}}$ を結ぶ m-測地線が直交することを意味する．すると任意の a, b による $z_i=g^{-1}(a+bx_i)$ で構成される $\boldsymbol{z}=(z_1,\cdots,z_n)$ との間に一般化ピタゴラスの定理 [3, 4]

$$d_\phi(\boldsymbol{y},\boldsymbol{z})=d_\phi(\boldsymbol{y},\hat{\boldsymbol{y}})+d_\phi(\hat{\boldsymbol{y}},\boldsymbol{z}) \tag{4}$$

が成り立ち，右辺第一項は本モデルで説明できない情報量，右辺第二項は本モデルで説明できる情報量，と解釈される．ここで特に $z_i=\bar{y}$ とすれば，線形モデルにおいて全変動の分解を表す式を自然に拡張したものになっているが，GLM においても null モデル $\bar{\boldsymbol{y}}$ からのデータ $\boldsymbol{y}$ のダイバージェンスを，推定値 $\hat{\boldsymbol{y}}$ を経由したダイバージェンスに分解していると捉えることができる．

4　まとめ

正準リンク関数という仮定のもとで Bregman ダイバージェンスの最小化という指針に沿って GLM における回帰母数を推定する方程式を導き，独立変数で貼られる e-平坦な空間への従属変数の m-射影という幾何構造を見た．線形モデルが GLM に拡張されることで，考える空間はユークリッド空間から双対平坦空間に拡張される．

モデルで説明できない情報量は，線形モデルにおいては残差平方和 (RSS) と呼ばれる量に，GLM では逸脱度 (deviance) と呼ばれる量に対応付けられる [1, 4] が，これらはダイバージェンスという観点からは統一的に理解される．これによって，例えば null モデルからのダイバージェンスの分解を考えることで，線形モデルでよく評価される決定係数が GLM の場合にも自然に拡張されうる．

参考文献

[1] Nelder, J. and Wedderburn, R., Generalized Linear Models, Journal of the Royal Statistical Society: Series A **135** (1972), 370–384.

[2] Banerjee, A., Merugu, S., Dhillon, I. S. and Ghosh, J., Clustering with Bregman divergences, Journal of Machine Learning Research **6** (2005), 1705–1749.

[3] Amari, S. and Nagaoka, H., Methods of Information Geometry, American Mathematical Society, 2001.

[4] Eguchi, S. and Komori, O., Minimum Divergence Methods in Stastical Machine Learning, Springer, 2022.

複数の高次元小標本データにおけるスパース次元削減手法の検証

長谷川 弘貴 [1], 矢田 和善 [1], 岡田 幸彦 [1], 國松 淳 [1]
[1] 筑波大学
e-mail : s2420513@u.tsukuba.ac.jp

1　はじめに

主成分分析（PCA）は，標本数が次元数より多いという前提が崩れると，固有値の推定が不安定になってしまう [1]．これは，複数データから共通主成分を算出できる共通主成分分析（CPCA）でも同様である [2]．単一の高次元小標本（以降 HDLSS）データに対しては，ノイズ掃き出し法 [1] などが提案されているが，複数の HDLSS データに対する手法は限られている．本研究は，複数の HDLSS データに対する手法の NR-CPCA[2] に，A-SPCA[3] の理論を応用した次元削減手法 A-SCPCA（Automatic - Sparse Common Principal Component Analysis）を検証する．

2　理論・シミュレーション設計

NR-CPCA[2] では，複数の HDLSS データに対する共通推定固有値 $\tilde{\lambda}_i$ と，それに対応する推定固有ベクトル $\tilde{h}_i$ を次のように定義している．

$$\tilde{\lambda}_i = \hat{\lambda}_i - \frac{\mathrm{tr}(S_D) - \sum_{s=1}^{i} \hat{\lambda}_s}{(n_1 + n_2 + \cdots + n_k - 1) - i}, \qquad \tilde{h}_i = \frac{X - \overline{X}}{\sqrt{(n_1 + n_2 + \cdots + n_k - 1)\tilde{\lambda}_i}} \hat{u}_i$$

$X_i \in \mathbb{R}^{d \times n_i} (i = 1, \cdots, k)$ の次元数を d，標本数を $n_i (n_i \ll d)$ とする．さらに $X = (X_1 \cdots X_k) \in \mathbb{R}^{d \times (\sum n_i)} (\sum n_i \ll d)$ とし，$\bar{X}$ は標本平均ベクトル $\bar{x}$ を $\sum n_i$ 個並べた行列 $\bar{X} = [\bar{x}, \cdots, \bar{x}]$ とする．また，双対共分散行列 S_D から計算される固有値を $\hat{\lambda}_i$，固有ベクトルを $\hat{u}_i$ とする．共通推定固有値 $\tilde{\lambda}_i$ とそれに対応する推定共通固有ベクトル $\tilde{h}_i$ の計算は，$i = \min(n_1 - 2, \cdots, n_k - 2, d)$ まで実施される．推定された固有ベクトル $\tilde{h}_i = (\tilde{h}_{i(1)}, \cdots \tilde{h}_{i(d)})^T$ の要素を，$\left|\tilde{h}_{oi(1)}\right| \geq \cdots \geq \left|\tilde{h}_{oi(d)}\right|$ のように絶対値の降順に並び替える．ここで整数 $\tilde{k}_i$ を，次の条件 (1) を満たすように見つける．

$$\sum_{s=1}^{\tilde{k}_i - 1} \tilde{h}^2_{oi(s)} < \omega_i \quad \& \quad \sum_{s=1}^{\tilde{k}_i} \tilde{h}^2_{oi(s)} \geq \omega_i \qquad (\omega_i \in (0, 1]) \tag{1}$$

ここで得られた整数 $\tilde{k}_i$ を用いて，固有ベクトル $\tilde{h}^*_i = (\tilde{h}_{i^*(1)}, \ldots, \tilde{h}_{i^*(d)})^T$ を次の式 (2) のように推定する．この方法により，各 i に対して $\tilde{h}^*_i$ を使用して i 番目の固有ベクトルを推定する．

$$\tilde{h}_{i^*(i')} = \begin{cases} \tilde{h}_{i(i')} & \text{if} \left|\tilde{h}_{i(i')}\right| \geq \left|\tilde{h}_{oi(\tilde{k}_i)}\right| \\ 0 & \text{else} \end{cases} \qquad (i' = 1, \ldots, d) \tag{2}$$

本研究のシミュレーションは，[3] の 1 つ目のシミュレーションに倣い，CPCA，NR-CPCA，A-SCPCA の 3 手法の比較を行う．設定固有値を $\lambda = (d^{2/3}, d^{1/2}, 1, \cdots, 1)$，対応する固有ベクトルを $h_1 = (1, 0, 0, \cdots, 0)^T, h_2 = (0, 1, 0, \cdots, 0)^T$，他の固有ベクトルは -1 から 1 の乱数から生成した物とする．共分散行列 Σ を用いて，d 次元多変量正規分布 $N_d(0, \Sigma)$ を定義する．用いるデータは，$X_1, X_2 \sim N_d(0, \Sigma)$ とする．このデータに対して，CPCA，NR-CPCA，A-SCPCA を実施し，平均二乗誤差 MSE で評価する．次元数 d は，$d = 2^s (s \in \{5, \cdots 11\})$，標本数は，$n_1 = n_2 = \frac{\lceil d^{1/2} \rceil}{2}$ とする．条件ごとにシミュレーションを 1000 回行い，その平均値を分析に用いる MSE として採用する．

3　結果

図 1 は，A-SCPCA(本研究では，$\omega_i = 1, \omega_i = 0.1$ の 2 パターンを採用)，NR-CPCA および CPCA を用いて実施したシミュレーションの結果である．図 1 によると，次元 d が増加するにつれて，すべての手法の MSE が減少し，手法間の MSE の差が小さくなっている事が確認された．A-SCPCA に着目すると，閾値ごとに結果に有意差があることが示された．特に $\omega_i = 0.1$ の場合，ほかのどの手法よりも有意に MSE が小さいことが示された（表 1)．

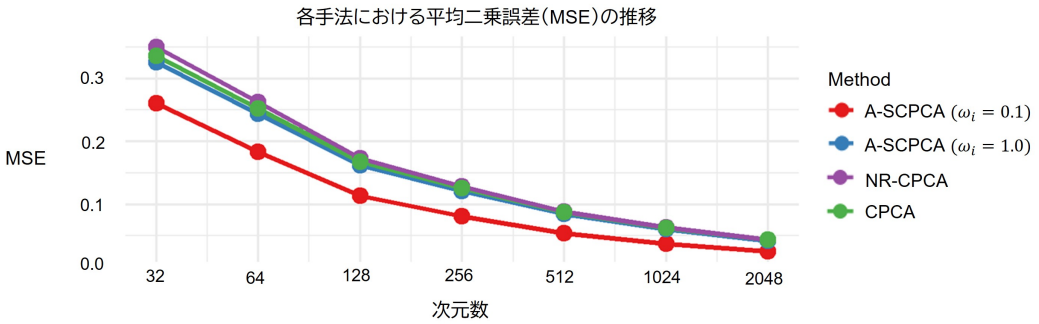

図 1. シミュレーションの結果

	A-SCPCA($\omega_i = 0.1$)	A-SCPCA($\omega_i = 1.0$)	NR-CPCA	CPCA
A-SCPCA($\omega_i = 0.1$)	–	0.0009742^{***}	0.002281^{***}	0.001332^{***}
A-SCPCA($\omega_i = 1.0$)	–	–	0.02324^{*}	0.009646^{**}
NR-CPCA	–	–	–	0.04436^{*}
CPCA	–	–	–	–

表 1. Welch-t 検定で得られた p 値の結果（$***: p < 0.005$,　$**: p < 0.01$,　$*: p < 0.05$）

4　結論

本研究は，複数の HDLSS データに対する次元削減手法 A-SCPCA を提案し、その有効性をシミュレーションにより検証した．その結果，A-SCPCA は，NR-CPCA および CPCA に比べて，MSE が有意に小さくなることが確認された．この結果は，閾値 ω_i を最適化する事で，より良い結果を示す可能性がある．今後は，実データへの適用を進め，その有効性を検証していく．

謝辞　本研究は，JST，さきがけ，JPMJPR21S4 の支援を受けたものである．

参考文献

[1] Yata, K. & Aoshima, M. Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. Journal of multivariate analysis. 105, 1, 193–215, 2012.

[2] Hasegawa, H., Kawamura H., Shin R,. Yata K., Okada Y. & Kunimatsu J. Noise Reduced Common PCA for High-Dimensional, Low-Sample Size Multi-View Data. in Proc. of The 6th International Conference on Statistics: Theory and Applications, 2024.

[3] Yata, K. & Aoshima, M. Automatic sparse PCA for high-dimensional data. Statistica Sinica, 35, 2025 (in press)

