（5）「バイオスタティスティックスの数理的基礎」に関する研究報告

ﾆｭｰﾗﾙﾈｯﾄ、生存時間解析、そしてﾘｻﾝﾌﾟﾘﾝｸﾞ法

大阪電気通信大学　総合情報学部　辻谷将明

1. はじめに

共変量の値が時間とともに変化する**時間依存型**データが含まれる場合、従来の Cox の比例ハザードモデルには、多くの問題点が存在する(Altman and Stavola, 1994)。Efron(1988)は、グループ化されたデータについて、**部分尤度**(Cox,1975)に基づく**部分**ロジスティック回帰を提案した。更に近年、ニューラルネットによる生存時間解析が、Cox の**比例ハザードモデル**に対抗して注目されつつある(Biganzoli et al., 1988,2002)。本報告では、グループ化されていないデータに関する 1)**部分**ロジスティックモデル、および 2)階層型ニューラルネットワークモデルによる生存時間解析を試みた。本手順の利点としては、1)時間依存型変数の取扱いが容易、2)観測期間全体を考慮に入れて解析、3)比例ハザード性の制約が不要、4)共変量の非線形性を考慮、5)死亡例が少ない場合も適用可などが挙げられる。

2. 部分ロジスティックモデル

本報告は、Mayo クリニックに来院した 312 例の PBC(原発性胆汁性肝硬変)患者を取上げた(Therneau et al., 2000)。最初は、6 ヶ月、12 ヶ月目、その後、1 年おきに来院する。そして、来院ごとに、年齢と共にプロトロンビン時間、ビリルビン値、アルブミン値、エデマ・スコアを測定する時間依存型データになる。目的は、任意の時点における 6 ヶ月後の生存率の推定にある。患者#d に関する時間依存型共変量を $\boldsymbol{X}_l^{<d>}=\left(t_l^{<d>},x_{l1}^{<d>},\cdots,x_{lI}^{<d>}\right)$ とする。ここに、$t_l^{<d>}$ は l 番目の時間区間の中央値である。Cox の比例ハザードモデルでは、相対ハザード $h(t)/h_0(t)$ が時間 t に依存するため、比例ハザード性が成立たない。例えば、表 1 は患者#9 の共変量の値を示している。1 人の患者が、7 個の観測値を生成していることになり、共変量の値が時間と共に変動する。

表 1 患者#9 の共変量の値

時間区間 l	中央値(日) $t_l^{<9>}$	年齢 (日) $x_{l1}^{<9>}$	ﾌﾟﾛﾄﾛﾝﾋﾞﾝ時間 $x_{l2}^{<9>}$	ﾋﾞﾘﾙﾋﾞﾝ値 $x_{l3}^{<9>}$	ｱﾙﾌﾞﾐﾝ値 $x_{l4}^{<9>}$	ｴﾃﾞﾏｽｺｱ $x_{l5}^{<9>}$
1	92.0	15526	11.0	3.2	3.08	0.0
2	272.5	15710	12.5	7.0	3.64	0.0
3	542.0	15887	11.2	4.2	3.10	0.0
4	875.0	16249	14.1	13.5	2.87	0.0
5	1211.5	16553	11.5	12.0	2.96	0.0
6	1837.0	16922	11.5	16.2	2.99	0.5
7	2339.0	17804	13.0	14.8	2.41	1.0

本報告の目的は、時間依存型共変量 $\boldsymbol{X}_l^{<d>}$ に基づいて、6 ヶ月後の生存時間を予測することにある。Efron(1988)は、グループ化されたデータについて、**部分尤度**(Cox,1975)に基づく**部分**ロジスティック回帰を提案した。これを拡張して、時間依存型共変量をもつグループ化されていないデータに対する部分ロジスティックモデル

$$h_l^{<d>}\left(\boldsymbol{x}_l^{<d>}\right)=\frac{1}{1+\exp\left[-\left\{\beta_0 t_l^{<d>}+\sum_{i=1}^{I}\beta_i x_{li}^{<d>}\right\}\right]},\ l=1,2,\cdots,l_d;d=1,2,\cdots,n \tag{1}$$

を提案した。

部分ロジスティックモデル(1)における共変量の有意性

$H_0:\beta_i=0, i=1,2,\ldots,I$

を検定するには、尤度比検定統計量

$$-2\ln\lambda = -2\left[l\left(\boldsymbol{X}_{[i]};\hat{\beta}'\right)-l\left(\boldsymbol{X};\hat{\beta}\right)\right], \tag{2}$$

を計算すればよい。ここに、$l\left(\boldsymbol{X};\hat{\beta}\right)$は初期標本$\boldsymbol{X}$に対する部分対数尤度であり、$l\left(\boldsymbol{X}_{[i]};\hat{\beta}'\right)$ は変量$\boldsymbol{X}^{<i>}$を除去した量である。帰無仮説が真なら、$-2\ln\lambda$は漸近的に自由度 $d.f.=1$ のカイ二乗分布に従う。また、離散時間に対する生存関数は

$$S(t_l)=\prod_{1\le i\le l}\left(1-h_i^{<d>}\right). \tag{3}$$

で与えられる。tまで生存した患者が、次の短期間Δt(例えば、6ヶ月)生存する条件付き確率は

$$\hat{\mathrm{P}}(t,\Delta t)=\frac{\hat{S}(t+\Delta t)}{\hat{S}(t)} \tag{4}$$

から推定される。

3. ニューラルネット

母集団構造の非線形性を摘出するだけでは不十分で、適切な非線形モデルで記述しなければならない。ニューラルネットワークモデルを想定し、ブートストラッピング(Ishiguro et al., 1997;Tsujitani et al.,2000)に基づくモデルの妥当性の検証、隠れユニット数の決定、および生存率の予測などについて考察した。

階層型ニューラルネットによって生存データを取扱う場合、ある入力パターンを提示したときの出力値は、ベイズの事後確率と考えられる。いま、$d\left(=1,2,\ldots,\sum_{d=1}^{n}l_d\right)$番目の入力$X_l^{<d>}=\left(t_l^{<d>},x_{l1}^{<d>},x_{l2}^{<d>},\ldots,x_{lI}^{<d>}\right)$が与えられたとき、最終出力(**離散ハザード率**)

$$h_l^{<d>}(\boldsymbol{x}_l^{<d>})=\frac{1}{1+\exp(-v_l^{<d>})}=\frac{1}{1+\exp\left[-\sum_{j=0}^{J}\left\{\frac{\beta_j}{1+\exp\left(-(\alpha a_l^{<d>}+\sum_{i=0}^{I}\alpha_{ij}x_{li}^{<d>})\right)}\right\}\right]} \tag{5}$$

が得られる。ブートストラッピングを用いて隠れユニット数を決定した。

参考文献

Altman, D.G. and De Stavola, B.L.(1994). ":Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates," *Statisics in Medicine*, 13, 301-341.

Biganzoli, E., Boracchi, P., and Marubini, E.(2002). "A general framework for neural network models on censored survival data," *IEEE Transactions on Neural Networks*, **15**, 209-218.

Biganzoli,E., Boracchi,P., Marriani,L. and Marubini, E. (1988). "Feed forward neural networks for the analysis of censored survival data: a partial logistic approach," *Statistics in Medicine*, **17**, 1169-1186.

Cox, D.R. (1975). "Partial likelihood," *Biometrika*, **62**, 269-276.

Efron, B.(1988). "Logistic regression, survival analysis, and Kaplan-Meier curve," *Journal of the American Statistical Association*, **83**, 414-425.

Ishiguro, M., Sakamoto, Y. and Kitagawa, G. (1997). "Bootstrapping log likelihood and EIC, an extension of AIC, " *The Annals of the Institute of Statistical Mathematics*, **49**, 411-434.

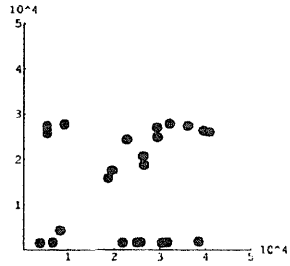
Therneau, T.M., Grambsch P.M.(2000). *Modeling Survival Data: Extending the Cox Model.* Springer: New York.

Tsujitani,M. and Koshimizu,T.(200) : Neural Discriminant Analysis.*IEEE Transactions on Neural Networks*, **11**, 1394-1401.

遺伝子型のモデルに基づく分類法の問題点

藤澤洋徳（統数研），江口真透（統数研），松浦正明（癌研ゲノムセンター），
宮田敏（癌研ゲノムセンター），牛嶋大（癌研ゲノムセンター）

1. はじめに



ゲノム・プロジェクトの進展に伴い，ハイスループットの遺伝子情報を短時間で処理できる実験系が開発されてきている．このような実験系の中でも，２次元平面に図のようにプロットされた個体のデータを基に，その位置関係から個々の個体における遺伝子型を判定する問題が存在する．

各ポイントは個体に対応し，座標の位置によって遺伝子型が示される．原点近傍は，実験を検査するサンプルと正常に反応が出なかったサンプルから構成される．Ｘの値が大きいとき，遺伝子型１を持っていると判断され，Ｙの値が大きいとき，遺伝子型２を持っていると判断される．Ｘ軸に近いサンプルは遺伝子型が１１であると判断され，Ｙ軸に近いサンプルは遺伝子型が２２であると判断され，傾き45度に近いサンプルは遺伝子型が１２であると判断される．

以下では，図のようなデータを基に各個体の遺伝子型をよりよく判断する方法について考察する．最初に，データの特徴を記述し，次に，一般のクラスタリング法の問題点を指摘し，モデルに基づく分類法を解説し，最後に改良点に関して議論する．

2. データの特徴

サンプルが比較的良く集まり肉眼でみてもグループを構成している事がはっきりわかる図のような典型的なデータだと，どのような方法でもうまくいくであろう．ところが，現実には多様なタイプが観測される．各グループのばらつきが，さらに大きい場合もあり，そのような場合は境界のサンプルの判別が問題となる．ばらつきも極端に異なる場合もある．原点近傍のグループは図では孤立しているが，例えば，遺伝子型１１のグループと連なっている場合もある．幾つかのグループが存在していない場合もある．グループに属するサンプルが１つという場合もある．外れ値が存在している場合もある．

このような状況の中で，はっきりと遺伝子型が分類できる個体はその遺伝子型を判定し，はっきりとは遺伝子型が分類できないような曖昧なサンプルについては判定を控え，「遺伝子型不明」とした結果をデータ解析者に返す

ことが必要となる．分類としてはミス・ジャッジを犯すことは許されず，判定できるものを正確に判定するルールが重要となる．分類が曖昧となるようなサンプルについては再度実験を行ない，正確な判定を得るまで保留できることに依っているためである．

3. モデルに基づくクラスタリング

図のようなデータをクラスタリングする方法として幾つか提案されている．本節では，一般的な方法を用いた場合のクラスタリングの問題点と，モデルに基づくクラスタリングの可能性を議論する．

代表的な統計的クラスタリングとしては k-means 法が使われている．誤ったクラスター数を設定すると，無理やりにグループを分割することもある．曖昧なサンプルがある場合，本来「遺伝子型不明」として判定すべきサンプルを無理やりどれかのクラスターに分類してしまう可能性もある．他のアドホックなクラスタリング方法も使われているが，同様な問題点が指摘できる．

モデルに基づくクラスタリングでは，各グループはある確率密度関数に従っていると考える．確率密度関数におけるパラメータ値は何らかの方法で設定することになる．それぞれのサンプルはベイズ規則によりクラスタリングができる．この場合，どのグループに属するかが曖昧なサンプルもベイズ事後確率で曖昧さを示唆できる．適当な信頼確率によって外れ値を指摘することも可能である．この方法を直接に適用するには，クラスター数の設定などの問題点がまだ残っているが，それらは次節で整理する．

4. データに応じて改良すべき点

前節で確率密度関数を想定した場合の標準的な適用法を説明したが，適用するデータの特徴を考えて，適用法を改良することが重要であろう．２次元のデータに直接にクラスタリング法を用いることは必ずしも効果的ではないかもしれない．遺伝子型グループと原点近傍グループが強く接触している場合には，判別できない．角度データが本質的に見えるが，始点が原点とは限らず，大きくずれる場合もある．各グループは極端なばらつきの違いを見せることもあるが，それがベイズ規則によるクラスタリング結果を奇妙に見せることがある．外れ値が存在するときには，パラメータの典型的な設定方法である最尤法は外れ値に引きずられやすく，分散の過大評価により，外れ値の同定やクラスタリングの誤りを起こしうる．あるグループのサンプルがあるかどうかは事前には分からず，場合によっては一つの場合もある．これらを克服する方法論が求められる．

グラフィカル・ガウシアン・モデリングの発現プロフィール解析への適用

堀本　勝久

東京大学医科学研究所ヒトゲノム解析センター
バイオスタティスティクス人材養成ユニット

1．序論

近年,一種の生物の全遺伝子数に相当する,数千もしくは数万の遺伝子の様々な環境下での発現量（発現プロフィール）が,マイクロアレイ技術の飛躍的な進歩によって計測できるようになった.現在,発現プロフィール解析には,大まかに二つの流れがある.一つは,発現プロフィールを用いて,類似な発現プロフィール・パターンを示す遺伝子群に分類する試みである.もう一つは, 発現プロフィールから遺伝子間の制御ネットワークを推定する試みである.我々は,標準的な統計解析法とグラフィカル・ガウシアン・モデリング（Graphical Gaussian Modeling: 以下 GGM）を組み合わせることで,発現プロフィールに基づいた遺伝子分類とネットワーク推定を同時におこなうことを試みた [1)-3)].

2．解析データ

本稿で扱う発現プロフィールのデータ形式を $p(i,j)$とし, 遺伝子 i の計測環境 j における発現量を表す. ここで, i は遺伝子の種類を表わし, 1 から N までとする. j は計測環境を表わし, 1 から M までとする. ここで計測環境とは, 細胞周期などに沿った時間的な計測環境や遺伝子が発現している組織や臓器の違いなど空間的な差異に基づく計測環境などがある.

3．発現プロフィールによる遺伝子分類

まず,発現プロフィールに対して階層的クラスター法を適用して,遺伝子の発現プロフィール間の類似性を階層的な形で求める. 2 つの遺伝子 i と j の発現プロフィール間の距離 d_{ij} は,以下の式のように定義した.

$$d_{ij} = \sqrt{\sum_{l=1}^{N} \left(r_{il} - r_{jl}\right)^2}$$

上式で,r_{il}は遺伝子 i と l との計測点に関する発現量間の Pearson の相関係数である.クラスター法は,群平均法を採用した.個々の遺伝子発現プロフィールを計測点数の次元から成るベクトルと考え,クラスター法によって得られたある階層のクラスター間で,個々の遺伝子の発現プロフィール・ベクトルの独立性を評価する.ベクトル間の独立性を見積もる方法として,我々は重回帰分析において説明変数間の多重共線性(multicolinearity)を評価する指標として用いられる variance inflation factor (VIR)を採用した.発現プロフィールへの応用に際しては,VIF が条件を満たすかどうかを,階層の低いレベルから順に判定していき,条件が満たされる最初の階層でのクラスター数を求める.

4．GGM の適用によるネットワーク推定

大量の発現プロフィールでは多くの場合,非常に類似した発現パターンを示すプロフィールが含まれる.すなわち,相関係数行列が極めて強い従属性を示すベクトルで構成されており,このような場合,相関係数行列の逆行列が数値計算上求められないことが起こる.そのため,実際の観測データに GGM を適用する際に採用される, Wermuth-Sheidt のアルゴリズムによる共分散選択を実行することが困難なる. そこで我々は,前節で解説したクラスタリングを導入し,この問題を回避した.推定されたクラスター数において,同一クラスター内の発現プロフィールについて,各計測点で発現量の平均値を求め,各クラスターを代表する発現プロフィールとする.これらの平均発現プロフィール間の相関係数行列について,GGM を適用した.階層的クラスタリングに沿って,発現プロフィール間の多重共線性の有無という基準を導入することで,この基準を満たすクラスター間の相関係数行列では,逆行列が求められることがほぼ保証されている.これによりクラスター数推定とネットワーク推定が,一つの方法に統合されている.

5．適用例

我々の方法を,Gasch らが酵母の 6152 の遺伝子について 173 の環境で計測した発現プロフィールに適用した[4].その結果,30 クラスターで VIF はすべて 10 以下になり,この結果,クラスター数として 30 が見積もれた.30 クラスターの構成遺伝子数の最大値と最小値は,727 と 6 であった.

次に GGM の適用した結果,435 の偏相関係数行列の要素の内,179 個の要素（約 41.1%）は 0 と見積もられ,残り 256 の要素（約 58.9%）は 0 ではないと見積もられた（図１）.推定されたネットワークにおいて,他のクラスターとのつながりが最も多く存在するクラスターでは,22 のクラスターとつながりが見積もれた.逆に,最もつながりが少ないクラスターでは,11 のクラスターとつながっている.またどのクラスターともつながりの存在しないクラスターは存在しなかった.

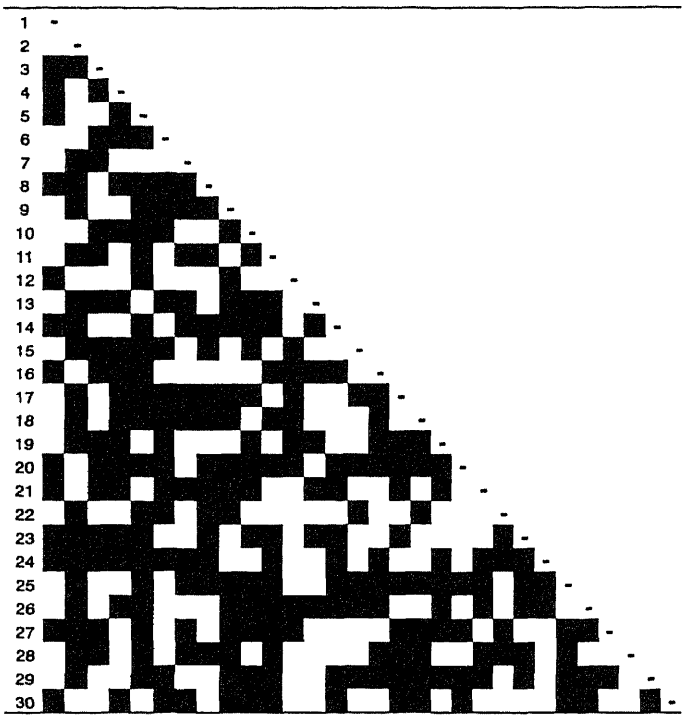


図１　推定されたネットワーク
白ヌキと黒ヌキは,それぞれ辺有り,辺無しを示す.

6．議論

我々の方法を要約すれば,階層的クラスタリングによるデンドログラムに沿って,クラスター数を自動的に推定し,それらクラスター間のネットワークを推定する.そして,適用例で示したように,大量遺伝子間の制御関係について大まかな枠組みを提供することができる.解析において設定を必要とするパラメータは,VIF の閾値（通常 10.0 を設定）と逸脱度の有意確率（5%を設定）のみである.今後の改良点としては,相関構造だけでなく因果関係を見積もる方法への改良や既知の生物学的知見を予め束縛条件として導入したモデルの開発などが挙げられる.

現方法では,遺伝子の１対１のネットワーク推定をおこなえない.今後,ベイジアン・ネットワーク法などによる遺伝子の１対１の推定法と我々の採用したクラスター対クラスター推定法のすり合わせによって,より広範囲かつ詳細なネットワーク推定が行なえることが期待される.また,本稿で解説した方法を,ASIAN　(Automatic System for Inferring A Network)というソフトウェア[5]として,公開予定である.

文献

1) Horimoto, K. and Toh, H. (2001) *Bioinformatics*, **17**, 1143-1151.
2) Toh, H. and Horimoto, K. (2002) *Bioinformatics*, **18**, 287-297.
3) Toh, H. and Horimoto, K. (2002) *J. Biol. Phys.* **28**, 449-464.
4) Aburatani, S., Kuhara, S., Toh, H. and Horimoto, K. (2002) *Signal Processing*, in press.
5) Horimoto, K., Aburatani, S., Kuhara, S. and Toh, H. (2002) *Res. Commun.Biochem Cell & Mol. Biol.*, in press.

A Monte Carlo simulation study on dimension reduction methods related to SIR

大瀧　慈 (広島大学原爆放射線医科学研究所)
Li-Xing Zhu (University of Hong Kong)

1. Principal Point-based SIR and Related Algorithms

The original version of SIR may suffer the difficulty of estimating the subspace of $\mathcal{B}$ where $E(\mathbf{z}|y)$ is a constant (or near a constant) function. To overcome this problem, we developed an algorithm PPSIR using the principal points notion (Forgy, 1965; Flury, 1990).

Another possible algorithm is a combination of SIR and SAVE as follows. Recall SAVE use $E(I_p - cov(\mathbf{z}|y))^2$ to find the directions. Based on the comments on the SIR and SAVE, we would simply consider a linear combination of SIR and SAVE to handle various cases. That is, rather than using each individual $cov(E(\mathbf{z}|y))$ or $E(I_p - cov(\mathbf{z}|y))^2$, we use the eigenvectors of $\mathbf{\Lambda} = \alpha\, cov(E(\mathbf{z}|y)) + (1-\alpha)\, E(I_p - cov(\mathbf{z}|y))^2$ to determine a subspace of $\mathcal{B}$. Denote by $\mathcal{B}_2$ the subspace determined by SAVE in terms of $(1-\alpha)E(I_p - cov(\mathbf{z}|y))^2$. The subspace determined by $\mathbf{\Lambda}$ is $\mathcal{B}_1 \cup \mathcal{B}_2$. With the estimation procedures of SIR and SAVE, we can have SIR-SAVE.

2. Monte Calro Study for Single Index Models

To assess the finite sample performance of the proposed algorithm, we conducted a Monte Carlo study. The performance measure used is the average of the correlation coefficient R^2 defined in Li (1991) from a total of 200 Monte Carlo sample. The following four models are considered.

$$\textbf{Model 1: } y = (\beta^T\mathbf{x})^2 \times \varepsilon;$$
$$\textbf{Model 2: } y = (\beta^T\mathbf{x})^2 + \varepsilon;$$
$$\textbf{Model 3: } y = \beta^T\mathbf{x} + \varepsilon;$$
$$\textbf{Model 4: } y = (\beta^T\mathbf{x})^3 + \varepsilon.$$

The covariable $\mathbf{x}$ and the error ε are taken to be independent normal $N(0, I_{10})$ and $N(0,1)$, where I_{10} is the 10×10 identity matrix. In performing the simulation, $\beta = (1, 0, \cdots, 0)$ and the sample size was 400. In the implimentation of PPSIR, we adopted $k = 2$ for the number of principal points for each slice, using the critical value $\delta =$

2.0, 3.0 and 4.0. To look into the sensitiveness of the algorithm for the number of slices, we considered $H = 2, 3, 5, 10, 20$.

From the simulation study we found the following: SAVE and pHd perform well for Model 2 with even regression function, $f(\beta^T\mathbf{x}) = (\beta^T\mathbf{x})^2$ but does not for Model 3 and 4 with odd function; For the model with the direction β in the error term, see Model 1, SAVE is also a good method. SIR can do well for Model 3 and 4. but it has difficulty to find the direction in even function and in the error term. These observations fit the justification described previously. Another observation is that SIR is quite insensitive to the choice of the number of slices. This has been illustrated by Li(1991) empirically and proved by Zhu and Ng (1995), while SAVE may be much more sensitive to the number of the slices. It can be seen more clearly for Model 4. Among these algorithms, PPSIR's performance is quite encouraging. It can find the direction for all five models and similar to SIR, is not sensitive to the slicing. It is however that due to the sensitiveness of SAVE to the number of slices, it does not maintain the advantage that SIR or PPSIR are not affected by choice of slice number significantly.

3. Monte Calro Study for Models with Two EDR Directions

For the models with two directions, we consider three models. Model 5 was used in Li(1991).

Model 5. $y = (\beta_1^T\mathbf{x}) \times (\beta_2^T\mathbf{x} + 1) + \varepsilon;$

Model 6. $y = (\beta_1^T\mathbf{x})^3 + (\beta_2^T\mathbf{x})^2 + \varepsilon;$

Model 7. $y = (\beta_1^T\mathbf{x})^2 + (\beta_2^T\mathbf{x})^2 \times \varepsilon.$

As we found SIR works well for Model 5, similar to the reported results in Li(1991), while not for the other models conducted here. Therefore, we report on the results with the other algorithms. For pHd, the directions cannot be estimated well with Model 5, 6 and for Model 7 the direction in the regression function can be found, but not for the one in the error term. SAVE performs slightly better than pHd for Model 5 and 6 in which both directions are in the regression function. For these models, the sensitiveness of SAVE to the slice number shows up again. As this model is in favor of SAVE, when H is relatively small, SAVE works well. As for PPSIR, its performance with Model 5 is similar to SIR, see Li(1991), and for the other models, PPSIR can only get one of the directions. Overall, SIR-SAVE algorithm enhances the ability for finding the directions. Again, it is sensitive to the choice of slice number due to the effect of SAVE.

サポートベクターマシーンによる識別・判別の理論と応用

九州大学大学院数理学府　安道 知寛[1]
九州大学大学院数理学研究院　小西 貞則[2]

超平面による特徴空間の分割に起源をもつ Support Vector Machine (SVM) (Vapnick 1998) は，文字認識などへの学習能力の高さから，近年様々な分野から注目を集めている．本報告では，様々な角度からSVMの性質を概観し，予測精度の高いSVMの構成法について検討した．特徴空間での kernel trick と soft margin 法を組み合わせることで訓練データを高精度で学習できる反面，margin に対する罰則パラメータ，及び kernel 関数の設定によっては過学習の問題が生じ，これらの選択がSVMの構築に当たっては本質的となる．VC次元 (Vapnick 1998) に基づき選択する方法も考えられるが，非常に粗い最悪評価であり，実際問題への適用上さらに精密な評価が望まれている．さらにVC次元は，使用可能な kernel 関数が制限されており，Gauss kernel などを用いた場合にはVC次元は利用できず，cross validation (CV) 法を優越する明白な利点は無いとの報告もある (Hastie *et.al* (2001))．

本報告では，reproducting kernel Hilbert space (RKHS) におけるSVMの損失関数を正則化理論に基づき明らかにし，CV法を精密化したブートストラップ法 (Efron & Tibsirani (1993)) によりモデル選択をおこなった．一般に用いられているブートストラップ法では，推定の変動が大きいことが指摘されているが，その推定変動を大幅に減少させる効率的リサンプリング法を提案し，シミュレーション回数を効果的に減らすことができた．

現在，SVMを種々の実際問題へ適用するにあたり，多群判別問題への展開や確率モデルの導入を試みる研究が盛んにおこなわれている (Lin *et.al* (2002), Tanabe (2001), Wahba *et.al* (2000), Zhu & Hastie (2001))．例えば，人顔認識は多群判別の典型例であり，セキュリティへの応用など現代社会の様々な場面での活用が考えられる．また，SVMに尤度の概念を導入することで，診断医療において病気の診断を確率的に下すことが出来たり，クレジットスコアリングにおいては信用リスクに応じたローン貸出が

[1] *Address for correspondence:* Tomohiro Ando, Graduate School of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan. E-mail: ando@math.kyushu-u.ac.jp
[2] E-mail: konishi@math.kyushu-u.ac.jp

可能となり，その応用の幅が一段と広がる．本報告では，多群判別確率モデルとしてSVMを定式化し，これら二つの要請を解決する手法を提案した．確率的SVM多群判別モデルの構成において本質的問題点は，正則化パラメータ及び kernel 関数の選択である．一般に，AIC (Akaike (1974)) と BIC (Schwarz (1978)) は最尤法に基づき構成したモデルの評価規準であるが，確率的SVM多群モデルは正則化法に基づいて推定していることから，当てはめたモデルの複雑さを適切に評価するモデル評価規準を新たに導出する必要がある．ここでは，AIC, BICを拡張し，確率的SVM多群モデルの汎化能力を評価するモデル選択規準を情報量，及びベイズ理論の異なる観点から提案した．データ数が多くなるに従い，モデルの候補が天文学的な数になってしまうため，ここでは遺伝アルゴリズム (Goldberg (1989), Holland (1975)) の使用を提案した．

提案したモデル及びモデル評価規準の有効性を，実データ及び人工データの解析を通して比較検討し，その結果，提案した手法は極めて有効な手法であることが分かった．

参考文献

Akaike, H. (1974): A new look at the statistical model identification. *IEEE Trans. Autom. Contr.* **AC-19** 716-723.

Efron, B. & Tibsirani, R. J. (1993): *An Introduction to Bootstrap.* Chapman & Hall.

Goldberg, D. E. (1989): *Genetic Algorithm in Search, Optimization and Machine Learning.* Addison - Wesley.

Holland, J.H. (1975): *Adaptation in Neural and Artificial Systems.* University of Michigan Press, Ann Arbor.

Kimeldoff, G. and Wahba, G. (1971): Some results on Tehebycheffian spline functions. *J. Math. Anal. Applic.* **33**, 82 -95.

Lin, Y., Wahba, G., Zhang, H., and Lee, Y. (2002): Statistical Properties and Adaptive Tuning of Support Vector Machines. *Machine Learning*, 48, 115 - 136.

Schwarz, G. (1978): Estimating the dimension of a model. *Ann. Statist.* **6**, 461 - 464.

Tanabe, S. (2001): Penalized Logistic Regression Machines: New methods for statistical prediction 1. *The ISM Cooperative Research Report*, 143, 163 - 194.

Vapnick, V. (1998): *Statistical learning theory*, New York, Wiley.

Wahba, G., Lin, Y. and Zhang, H. (2000) Generalized Approximate Cross Validation for Support Vector Machines, or, Another Way to Look at Margin-Like Quantities. in *Advances in Large Margin Classifiers*, MIT Press, 297 - 309.

Zhu, J. & Hastie, T. (2001): Kernel Logistic Regression and the Import Vector Machine. NIPS2001 conference.

多変量ノンパラメトリック回帰におけるバイアス縮小推定量について

島根大学　総合理工学研究科　　吉崎正浩
島根大学　　総合理工学部　　内藤貫太

1. はじめに デザインが多次元（d 次元）の回帰関数を推定するノンパラメトリックな方法として Kernel を用いる局所線形推定量はよく用いられ，その性質も盛んに議論されている [Ruppert and Wand(1994) など]．しかし，この方法は曲率が大きい付近の点や境界付近の点でのバイアスが目立つ事が知られている．そこで本稿では多項式の次数を増やすことなく，従来の局所多項式推定量を初期推定量と見なし，加法的な調整項を加えることによってバイアスの縮小を達成するという方法で得られた推定量 (ABRE) について考察を与える．

2. 推定量の構成 サンプルを $(X_1, Y_1), \cdots, (X_n, Y_n)$ とし，$X_i = (X_{i1}, \cdots, X_{id})^T$ とする．回帰モデル

$$Y_i = m(X_i) + v(X_i)^{1/2}\varepsilon_i$$

について考える．ただし，ε_i は i.i.d. で平均 0，分散 1 とする．$K(u) = K(u_1, \cdots, u_d) \geq 0$ は Kernel 関数，$H^{1/2}$ は Bandwidth Matrix であり，$K_H(u) = |H|^{-1/2}K(H^{-1/2}u)$ として，

$$X_{1,x} = \begin{pmatrix} 1 & (x - X_1)^T \\ \vdots & \vdots \\ 1 & (x - X_n)^T \end{pmatrix}, \ X_{2,x} = \begin{pmatrix} 1 & (x - X_1)^T & \mathrm{vech}^T[(x - X_1)(x - X_1)^T] \\ \vdots & \vdots & \vdots \\ 1 & (x - X_n)^T & \mathrm{vech}^T[(x - X_n)(x - X_n)^T] \end{pmatrix},$$

$$W_x = \mathrm{diag}\{K_H(x - X_1), \cdots, K_H(x - X_n)\}, \ Y = (Y_1, \cdots, Y_n)^T$$

とする時，多変量の Local Linear 推定量 $\tilde{m}_1$ は

$$\tilde{m}_1(x) = \boldsymbol{e}_1^T(X_{1,x}^T W_x X_{1,x})^{-1} X_{1,x}^T W_x Y$$

で与えられる．ただし，$\mathrm{vech}[A]$ は A の 1 列目から順番に対角成分以下を縦に並べたベクトルで，$\boldsymbol{e}_r$ は第 r 成分だけ 1 で他は 0 を持つベクトルである．$\tilde{m}_1$ を初期推定量と見なし，$\tilde{m}$ を $\tilde{m} + \xi$ という形で調整する事を考える．調整項 ξ は

$$L(\xi_0, \boldsymbol{\xi_1}^T|x) = \sum_{i=1}^{n} \left[Y_i - \tilde{m}_1(X_i) - \xi_0 - \boldsymbol{\xi_1}^T(x - X_i)\right]^2 K_H(x - X_i)$$

と最小化した時の $\tilde{\xi}_0$ とする．つまり調整項 $\xi = \xi(x)$ は "Local Linear 推定量の残差の Local Linear 推定量" となる．そして提案する推定量 (ABRE) は

$$\hat{m}_1(x) = \tilde{m}_1(x) + \tilde{\xi}(x) = \boldsymbol{e}_1^T(X_{1,x}^T W_x X_{1,x})^{-1} X_{1,x}^T W_x (2Y - M_1)$$

となる．ここで $M_1 = (\tilde{m}_1(X_1), \cdots, \tilde{m}_1(X_n))^T$ である．同様にして Local Quadratic 推定量を $\tilde{m}_2$ とした時の ABRE は，

$$\hat{m}_2(x) = \boldsymbol{e}_1^T(X_{2,x}^T W_x X_{2,x})^{-1} X_{2,x}^T W_x (2Y - M_2)$$

とできる．ここで $M_2 = (\tilde{m}_2(X_1), \cdots, \tilde{m}_2(X_n))^T$ である．

3. 推定量の挙動 提案した推定量は次の性質を持つ．ただし，$\mathrm{Bias}[\cdot|X_1,\cdots,X_n]$, $\mathrm{Var}[\cdot|X_1,\cdots,X_n]$ をそれぞれ $\mathrm{Bias}[\cdot]$, $\mathrm{Var}[\cdot]$ と表記する．

定理 x を $\mathrm{supp}(f)$ の内点とする．適当な条件のもとで，

$$\begin{aligned}\mathrm{Bias}[\widehat{m}_1(x)] &= -\frac{1}{4}\mu_2(K)^2\mathrm{trace}\left[H\nabla^2\mathrm{trace}\left[H\nabla^2 m(x)\right]\right]+o_p(\mathrm{trace}[H^2]),\\ \mathrm{Var}[\widehat{m}_1(x)] &= \frac{v(x)R(K^*)}{n|H|^{\frac{1}{2}}f(x)}+o_p(n^{-1}|H|^{-\frac{1}{2}})\end{aligned}$$

となる．ただし，$\nabla^2 m(x)$ は $m(x)$ のヘッセ行列で，

$$K^*(u)=2K(u)-K*K(u)\ ,\ K*K(u)=\int K(\boldsymbol{v})K(\boldsymbol{u}-\boldsymbol{v})d\boldsymbol{v}\ ,\ R(K^*)=\int K^*(u)^2du.$$

境界付近でも内点と同じオーダーを達成する．$\widetilde{m}_1$ は $O_p(\mathrm{trace}[H]), O_p(n^{-1}|H|^{-1/2})$ という漸近バイアスと漸近分散を持つこと [Ruppert and Wand(1994)] と比較すれば，漸近分散は同じオーダーで且つ漸近バイアスが小さくなっていることがわかる．また，同様にして $\widehat{m}_2$ の漸近バイアスは内点では $O_p(\mathrm{trace}[H^3])$ で，境界付近の点では $O_p(\mathrm{trace}[H^2])$ となることが示される．$\widetilde{m}_2$ の漸近バイアスは内点，境界付近の点でそれぞれ $O_p(\mathrm{trace}[H^{3/2}]), O_p(\mathrm{trace}[H])$ であり，$\widehat{m}_2$ と $\widetilde{m}_2$ は同じオーダーの漸近分散を持つことから，$\widehat{m}_2$ はバイアス縮小推定量であることがわかる．

4.Bandwidth の選択 ここでは Yang and Tschernig(1999) の方法を参考に $\widehat{m}_1$ で $d=2, H^{1/2}=hI_2, v(x)=\sigma^2$ とした場合のプラグインを用いた Bandwidth 選択に関して考える．ここで提案する方法は

$$\widehat{h}_{\mathrm{PI}}=h_{\mathrm{AMISE}}\left[1+O_p(n^{-1/6})\right]$$

を得る．

5. おわりに 本稿では従来用いられてきた局所多項式推定量を初期推定量とし，加法的調整項を加えることでバイアスを縮小する推定量について議論した．この加法的調整は X_x のサイズを変えることなくバイアス縮小を達成できるので，多項式をあてはめる次数を大きくしてバイアスを縮小することと比較すると，計算コストの面においても有効であると言える．Bandwidth の推定法の詳細，その Bandwidth や推定量に関するシミュレーション，実データへの適用に関しては当日報告する．

参考文献

1. Choi, E., Hall, P. and Rousson, V. (2000). Data sharpening methods for bias reduction in nonparametric regression. *Ann. Statist.* **28**, 1339-1355.

2. Naito, K. and Yoshizaki, M.(2002). Nonparametric Regression with Additive Adjustment. 統計数理研究所共同研究レポート 155, 115-128.

3. Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346-1370.

4. Yang, L. and Tschernig, R. (1999). Multivariate bandwidth selection for local linear regression. *J. R. Statist. Soc. B.* **61**, 793-815.

連絡先:s97499@math.shimane-u.ac.jp（吉崎正浩），naito@math.shimane-u.ac.jp（内藤貫太）

2項回帰問題におけるkernel平滑化とその生物検定法への応用

中国短大　奥村英則
島根大　内藤貫太

1. はじめに　2項回帰関数に対するノンパラメトリック推定について考える. K 個の共変量 $x_1, \ldots, x_K$ の各共変量 x_i で N_i 個体中反応した個体の数 Y_i が観測されたとする. ここですべての個体の反応は独立であるとする. 反応した固体数 Y_i, $i=1,\ldots,K$ はパラメータ $p_i=p(x_i)$ をもつ2項分布 $Bi(N_i,p_i)$ に従うとする. 回帰関数 $p(x)$ に関しては何も知見がない場合にはノンパラメトリックアプローチが使用される. 2値反応に対して, Copas(1983) が最初に kernel 推定量の1つである Nadaraya-Watson 推定量を使用した. このデータに対する Nadaraya-Watson 推定量は $\hat{p}_{NW}(x;h)=(\sum_{i=1}^K Y_i w_i)/(\sum_{i=1}^K N_i w_i)$ で与えられる. ここで, $w_i=\phi((x_i-x)/h)/h$ であって, $\phi(x)$ は原点対称な密度関数である. 本報告では, 共変量間での分散不均一性を考慮した kernel 推定量を提案し, その推定量の挙動を明らかにする. そして生物検定法への応用として, $p(x)$ の推定の逆問題を考える. その逆問題の解の推定量を我々の提案した $p(x)$ の推定量から構成し, その推定量の挙動について考察を与える. ここで $x_i=(i-1)/(K-1)$, $i=1,\ldots,K$ とする.

2. 重み付き推定量　データ $(x_i,\bar{Y}_i)$, $i=1,\ldots,K$ を kernel を用いて平滑化することを考える. ここで, $\bar{Y}_i=Y_i/N_i$, $i=1,\ldots,K$ である. このデータに対する Nadaraya-Watson 推定量 $\hat{p}_M(x;h)=(\sum_{i=1}^K \bar{Y}_i w_i)/(\sum_{i=1}^K w_i)$ を Staniswalis and Cooper(1988) が提案している. これは局所的に最小2乗法を適用するこによって導出できる. 確率変数 $\bar{Y}_i$ は, 分散 v_i/N_i を持つ. ここで $v_i=p_i(1-p_i)$ である. そこで各用量 x_i の分散 v_i/N_i の大きさに応じて, 重み w_i を変動させることを考える. この推定に対する1つのアプローチとして局所的に"重み付き最小2乗法"を適用することが考えられる. これによって導出される推定量は

$$p^*(x;h)=(\textstyle\sum_{i=1}^K Y_i w_i/v_i)/(\sum_{i=1}^K N_i w_i/v_i)$$

で与えられる. ここで v_i は未知であり, その意味で $p^*(x;h)$ は理想的な推定量である. 実際には v_i の推定量が必要である. p_i の推定量として $\hat{p}_i=(Y_i+\sqrt{N_i}/2)/(N_i+\sqrt{N_i})$, $i=1,\ldots,K$ を採用する. $\hat{p}_i$ は2項分布 $Bi(N_i,p_i)$ の p_i が無情報事前分布に従うと仮定したときのベイズ推定量として知られている. これを用いて v_i の推定量 $\hat{v}_i=\hat{p}_i(1-\hat{p}_i)$ が得られる. 最終的に, $p^*(x;h)$ の式の v_i に $\hat{v}_i$ を代入して得られる

$$\hat{p}(x;h)=(\textstyle\sum_{i=1}^K Y_i w_i/\hat{v}_i)/(\sum_{i=1}^K N_i w_i/\hat{v}_i)$$

が提案する推定量である. 基準 (1) は, $\bar{Y}_i$ は漸近的に平均 p_i, 分散 v_i/N_i の正規分布に従うので, 尤度原理によって支持されることに注意する.

3. 推定量の性質　共変量のレベルの個数 K を固定した場合と K が十分大きい場合において, 提案された推定量の挙動を MSE によって評価する. また, K が十分大きい状況での回帰関数 $p(x)$ の信頼区間を与える.

まず, 共変量のレベルの個数 K が固定されているときを考える. 推定量 $\hat{p}_{NW}(x;h)$, $\hat{p}_M(x;h)$ と $p^*(x;h)$ の MSE は容易に正確に計算できる. しかし, 提案した推定量 $\hat{p}(x;h)$ は, v_i の推定量を使用しているので MSE を正確に計算できない. そこで, K と h は固定し, $N=\sum_{i=1}^K N_i\to\infty$, $N_i/N\to 1/K(i=1,\ldots,K)$ とする状況を考える. このとき, $\hat{p}(x;h)$ の MSE は次のような形で書ける.

$$\mathrm{MSE}[\hat{p}(x;h)]=(\mathrm{Bias}[p^*(x;h)])^2+2\left\{\frac{G_1}{\sqrt{N}}+\frac{G_2}{N}\right\}\mathrm{Bias}[p^*(x;h)]+\frac{G_1^2}{N}+\mathrm{Var}[p^*(x;h)]+\frac{H}{N}+o(\tfrac{1}{N}).$$

ここで, G_1, G_2 は $\mathrm{Bias}[\hat{p}(x;h)]$ から, H は $\mathrm{Var}[\hat{p}(x;h)]$ から導出される.

次に, 共変量のレベルの個数が十分に大きい場合について考える. これは通常のノンパラメトリック回帰で議論される状況に対応する. 適当な条件のもとで次式が成り立つ.

$$\mathrm{MSE}[p^*(x;h)]=\mathrm{AMSE}[p^*(x;h)]+O\left(hK^{-1}+h^6\right),$$

$$\mathrm{AMSE}[p^*(x;h)] = v(x)R(N_1Kh)^{-1} + 4^{-1}h^4\mu_2^2\left\{p^{(2)}(x) - 2(1-2p(x))p^{(1)}(x)^2v(x)^{-1}\right\}^2.$$

ここで $R = \int_{-1}^{1}\phi(z)^2dz, \mu_2 = \int_{-1}^{1}z^2\phi(z)dz$ である. 理想的な推定量 $p^*(x;h)$ と Nadaraya-Watson 推定量 $\hat{p}_{NW}(x;h)$ との漸近的な MSE の違いは, それらの分散は等しいので, $O(h^2)$ バイアスに現れる. ある x に対して, $\mathrm{ABias}[p^*(x;h)]$ の 2 乗と $\mathrm{ABias}[p_{NW}(x;h)]$ の 2 乗の差から導出される式 $(1-2p(x))\{v(x)p^{(2)}(x) - (1-2p(x))p^{(1)}(x)^2\}$ が正であれば, $\mathrm{AMSE}[p^*(x;h)] < \mathrm{AMSE}[\hat{p}_{NW}(x;h)]$ が成り立つ.

$\hat{p}(x;h)$ のバイアスの式

$$\mathrm{Bias}[\hat{p}(x;h)] = \mathrm{ABias}[\hat{p}^*(x;h)] - (1-2p(x))N_1^{-1} + o\left(h^2 + N_1^{-1}\right)$$

に注目して $\tilde{p}(x;h) = N_1^{-1} + (1-2N_1^{-1})\hat{p}(x;h)$ と定義すれば, これは提案した推定量 $\hat{p}(x;h)$ のバイアス修正推定量になる. 適当な条件の下で,

$$\sqrt{KN_1h}(\tilde{p}(x;h) - p(x)) \longrightarrow_d N(\rho f(x)\mu_2, p(x)(1-p(x))R).$$

ここで, $f(x) = \{p^{(2)}(x) - 2(1-2p(x))p^{(1)}(x)^2v(x)^{-1}\}/2$ で, $\rho(\geq 0)$ は $KN_1h^5 \to \rho^2$ を満たすある定数である. 適当な kernel ϕ を使用すれば $\tilde{p}(x;h)$ の 1 次導関数 $\tilde{p}^{(1)}(x;h)$ と 2 次導関数 $\tilde{p}^{(2)}(x;h)$ はそれぞれ $p^{(1)}(x)$ と $p^{(2)}(x)$ の一致推定量となる. 故に, この kernel を使用することによって $p(x)$ の近似 $100(1-\beta)\%$ 信頼区間が次のように得られる.

$$\left[\tilde{p}(x;h) - h^2\mu_2\tilde{f}(x) - \Phi^{-1}(1-\tfrac{\beta}{2})V,\ \tilde{p}(x;h) - h^2\mu_2\tilde{f}(x) + \Phi^{-1}(1-\tfrac{\beta}{2})V\right]$$

4. 逆問題 (生物検定法への応用) 回帰関数 $p(x)$ が単調であるとき, $\alpha(0<\alpha<1)$ に対して, $\Theta_\alpha = p^{-1}(\alpha)$ に関する推定が必要とされる場合がある. 例えば, $p^{-1}(\alpha)$ は薬理学では, 薬物の $100\alpha\%$ 有効量であり 毒物学では毒物による $100\alpha\%$ 致死量にあたる. 生物検定法では, Θ_α の推定や区間推定などが行なわれる. Müller and Schmitt(1988) と同様にして, Θ_α の推定量を $\tilde{\Theta}_\alpha = (\inf M_\alpha + \sup M_\alpha)/2$ で定義する. 適当な条件の下で,

$$\sqrt{KN_1h}(\tilde{\Theta}_\alpha - \Theta_\alpha) \longrightarrow_d N(\rho f(\Theta_\alpha)\mu_2 p^{(1)}(\Theta_\alpha)^{-1}, \alpha(1-\alpha)Rp^{(1)}(\Theta_\alpha)^{-2})$$

が成り立つ. 故に, Θ_α の近似 $100(1-\beta)\%$ 信頼区間が次のように得られる.

$$[\tilde{\Theta}_\alpha - \{h^2\mu_2\tilde{f}(\tilde{\Theta}_\alpha) + \Phi^{-1}(1-\tfrac{\beta}{2})\sqrt{\tfrac{\alpha(1-\alpha)}{KN_1h}}\}\tilde{p}^{(1)}(\tilde{\Theta}_\alpha)^{-1},\ \tilde{\Theta}_\alpha - \{h^2\mu_2\tilde{f}(\tilde{\Theta}_\alpha) - \Phi^{-1}(1-\tfrac{\beta}{2})\sqrt{\tfrac{\alpha(1-\alpha)}{KN_1h}}\}\tilde{p}^{(1)}(\tilde{\Theta}_\alpha)^{-1}]$$

5. シミュレーション Θ_α の提案した推定量, Nadaraya-Watson 推定量と最尤推定量との比較をシミュレーションによって行なった. 比較は MSE, 推定量が計算不能な個数, 平均信頼区間, coverage probability によって行なった. 回帰関数 $p(x)$ にタイプ III の一般化ロジスティック分布関数や混合ロジスティック分布関数を使用し, いくつかの α に対してこれらの値を求めて比較を行った. Θ_α の最尤推定量は回帰関数がロジスティックモデルに属すると仮定し, 最尤法によって算出した. 我々の提案した推定量は, $\alpha = 0.5$ のとき点推定で, $\alpha = 0.1$ のとき区間推定で良い挙動を示した. それらの結果の詳細ついては実データへの適用と合わせて当日報告する.

6. おわりに Nadaraya-Watson 推定量は, 重み付けなしの局所的最小 2 乗推定量であって, 個々のデータとその平均との差の 2 乗和を基準として構成される. 一方, 我々の提案した推定量は, 分散の情報を考慮した局所的重み付き最小 2 乗推定量であって, それは局所対数尤度推定量と見ることができ, 対数尤度の意味でよい. それは直感的に誤差を標準化することの妥当性を裏付ける. 一般に Kernel に対する重み付けによる違いは漸近的には $O(h^2)$ バイアスのみに現れる. しかし, シミュレーションよる推定量の比較において, 我々の提案した推定量は良い挙動を示した. 従って, 2 項データにおける我々の行なった重み付けは有効であるといえる. データに基づくバンド幅 h の選択については今後の課題にしたい.

参考文献

Copas, J. B. (1983). Plotting p against x. *Applied Statistics* **32**, 25-31.

Müller, H.-G. and Schmitt, T. (1988). Kernel and probit estimates in Quantal Bioassay. *J. Amer. Statist. Assoc.*, **83**, 750-759.

Staniswalis, J. G. and Cooper, V. (1988). Kernel estimates of dose response. *Biometrics*, **44**, 1103-1119.

統計的検定の困難を回避する柔軟な試験計画

統計数理研究所　柳本 武美

1. 序

統計的検定は、科学的な知識を検証する基準として広く用いられている。一方で素朴な研究者の直観との喰い違いも見られる。対策としては、イ）素朴な認識論を改める（例えば 柳本，印刷中）、ロ）工夫したデータ解析を行う、ハ）試験計画に工夫を加える、が考えられる。一昔前迄は専らデータ解析法の研究に関心が集中していたが、近年は試験計画の工夫の重要性が理解されてきた。試験の計画と得られたデータの解析は独立した２つの仕事ではなく、一体のものとの理解が進んだためである。

この報告では、２つの良く知られた例について考察を行う。対象とする問題は普遍的であるが、議論は臨床試験を想定して行う。

2. 実際の水準の減少

２項母集団の出現確率 p についての、$H_0 : p = p_0$ の検定、あるいは２つの２項母集団の出現確率 p_1, p_2 についての、$H_0 : p_1 = p_2$ の検定を考える。良く知られているように、名目の水準 α (例えば0.05) より実際の水準が小さくなる。検定統計量が離散的であれば、この現象は避けられない。しかし実際の水準が $\alpha/2$ より小さくなることもまれではない。

2-1. 事後解析

データは所与であり、試験の計画に全く関与しないと仮定した場合を考える。この場合には既存の方法としては、イ）理論的帰結として受け入れる、ロ）Randomized 検定を行う、ハ）mid-p 検定を行う、がある。ハ）の mid-p 検定は、Frank (1986)、Barnard (1989) が議論しているが、科学的な推論としては意味不詳である。Randomized 検定も、同質の試験が繰り返して行われるのでなければ、採用され難い。

2-2. 延長戦方式

実際の水準の減少を回避することを試験計画に組み込むとすれば多くが考えられる。１つの方法は、検定統計量が棄却域に最近の値になった時またその時のみに、標本を追加して解析を行う。これは２段階逐次検定となる。逐次検定の性能の良さは広く知られているが、実際に数値計算を行って確かめることができる。この場合のポイントは、標本の追加を試験の開始前に定めておくことである。

3. 有意でない試験での標本追加

Cornfield (1966) が初めて明確に指摘したとされている。ある研究者が試験を実施して検定を行ったが有意ではなかった。そこで追加標本の大きさを決めようとしたが答えはない。もしその研究者が有意な結果を得ていれば、そこで試験を終えて結果を公表すると想定する。この場合有意でない結果を得た時にはその時点で止めるしかない。

3-1. 事後解析

多くの研究者は、追加試験は許容されると考えていると思われる。しかし推論が不明瞭であることは明白である。勿論実際に追加試験が行われた場合には推論の検討の余地がある。別の研究者が同じ試験計画書で試験を実施した場合と同じになる。一方で、事後解析で努力するより、試験計画に精力を注いだ方が研究者にとって有利なことも疑いがない。

3-2. 意図しない中止

上記の事実は、治療の効果を示したい研究者は大きい標本サイズの試験を計画しなくてはならないことを示している。ところで臨床試験では大きい標本サイズが望ましくない事情が二点ある。一点は非常に小さい治療効果は検出する必要がないことであり、他の一点は非常に小さい p-値は望ましくないことである。純粋科学の分野では、小さい p-値は、資源の浪費を無視すれば、一般的には望ましい。しかし臨床試験では、中間解析で p-値がある程度小さいと試験は中止すべきである。これは倫理的側面であって、論理的側面とは異なる。

もしも十分に大きいサイズの試験を行うに足る仮説でもあるならば、研究者は実際に試験を実施する。そして、第３者（例えば効果安全性評価委員会）は、倫理的に継続が認められ試験を中止する。これは研究者の本来の意図とは異なる中止となる。研究者は、意図しない中止となった場合には、関連した試験（投与量の変更、対象患者の拡大）を行う。第３者が中止を勧告しない場合には試験は継続されるから、Cornfield が指摘した困難（fallacy）は回避される。

Barnard, G.A.(1989). On alleged gains in power from lower P-values. *Statist. Med.*, **8**, 1469-1477.

Cornfield, J.(1966). Sequential trials, sequential analysis and the likelihood principle. *Ann. Statist.*, **29**, 18-23.

Frank, W.E.(1986). P-values for discrete statistics. *Biometrical J.*, **28**, 403-406.

柳本 武美. 実証主義の転換から導かれる統計的推論の視点, 日本統計学会誌, **32** 巻, 291-302.

遺伝統計学の基礎と応用

東京女子医科大学膠原病リウマチ痛風センター
大学院先端生命医科学系専攻遺伝子医学分野
鎌谷直之

1 遺伝統計学的な解析手法

遺伝統計学にはさまざまな手法があるが、メンデルの法則とその例外（遺伝法則）を直接用いた手法（連鎖解析）と間接的に用いた手法（連鎖不平衡を利用した解析）がある。それらの特徴を以下に示した。

表1　種々の連鎖解析、連鎖不平衡を利用した解析

解析の方法	適する対象集団	集団の構造化による問題	陽性領域
パラメトリック連鎖解析[1]	少数大家系	最小	広い
ノンパラメトリック連鎖解析[1]	多数小家系	小	広い
TDT、S-TDT	多数小家系	小	狭い
連鎖不平衡解析[2]	多数の患者と対象者	中	狭い
相関分析	多数の患者と対象者	大	狭い
家系内相関分析	多数の患者と家族	小	狭い

2 パラメトリック連鎖解析のアルゴリズムの基礎

パラメトリック連鎖解析の基礎を考える。

家系図、個人の連鎖した複数の座位での遺伝子型、個人の表現型、マーカー座位間の組み換え割合、位置不明の一つの疾患座位、疾患座位での遺伝子型毎の浸透率が与えられた時どのような解析が可能であろうか。

確率論に従って、一つの実験を考え、それにより可能なすべての結果（outcome）の集合を標本空間（Ω）とし、その部分集合を出来事 (event) とする. メンデルの法則に従い出来事に確率を後に述べるように対応させる. 標本空間は有限である.

このようなモデルに基づいて観察データである各座位の遺伝子型と個人の表現型の生じる確率、即ち尤度を求めると極めて多い項数の多項式ができる。しかしほとんどの項では0である。従って、0となる項をまとめて、共通部分を括りだすことができれば計算は速くなる。Elston-Stewart アルゴリズムでは、家系の最も子孫の核家族（両親とその子供すべて）について、データが矛盾する項を取り除き、残った項のみを計算するという手法を用いる。実際には、子の遺伝子型から親の可能なディプロタイプ形をリストアップする。

最も子孫の核家族の一つについて、先祖から伝わる配偶子伝達によるアレル伝達と組換え、新た

[1]狭義の連鎖解析
[2]ハプロタイプ解析を含む

な二つのハプロタイプの一つを選ぶ過程について、すべての出来事を $E_1, ..., E_Q$ とする。尤度関数の各項には必ず $Pr(E_1), ..., Pr(E_Q)$ のどれかが含まれているので、すべての項を Q 種類にまとめて $Pr(E_1), ..., Pr(E_Q)$ のいずれかを括りだすことができる。$Pr(E_1), ..., Pr(E_Q)$ の多くは 0 であり、それらの項を消すことができる。次に、残った項について、さらに上の核家族について同様な操作を行い（即ち、可能な出来事のみを残し不可能な出来事を消し去る)、最終的に一番上の核家族に到達する。

以上の計算では、核家族の出来事について多項式の複数の項をまとめ、多くの項を消し去るという操作を行うことがポイントである。

連鎖解析の手法としては、以上の方法の他に継承ベクトルを用いた Lander-Green アルゴリズムがある。

3 ハプロタイプ推定 LDSUPPORT とプールされたハプロタイプ推定 LDPOOLED のアルゴリズム

ハプロタイプ推定の問題を詳細に検討すると、プール問題に還元できる。一人の個人は二つのハプロタイプを持ち、その二つが混合していて情報が劣化している。そのような不完全な情報から完全情報を推定する、という問題がハプロタイプ推定問題の本質である。これを一般化すると次のような問題になる。

多くのハプロタイプが、それぞれ $2m$ 個のハプロタイプを含む多くの小集団に混合されている。そのため、それぞれの混合されたプール単位で劣化された情報が観察される。そのように、不完全になった情報から完全情報を推定する。このように、一般のハプロタイプ推定問題を拡張したプログラムが ldpooled である。通常のハプロタイプ推定の問題は、このように拡張された $2m$ 個のハプロタイププールからの推定の問題の中で単に $m = 1$ の特殊な場合である。

即ち、ldpooled で最大化される尤度関数は、

$$L(\Theta) = \prod_{k=1}^{n} \sum_{Q_{kj} \in R_k} \prod_{i \in Q'_{kj}} \theta_i \tag{1}$$

ただし、R_k は k 番目のプールのデータに合致するすべてのハプロタイプの順列の集合、Q_{kj} は R_k の j 番目の要素であるハプロタイプの順列、Q'_{kj} は Q_{kj} の因子であるハプロタプの番号の集合である。

確率的ネットワークアルゴリズム による継承分布の計算法に付いて

東京工業大学 情報理工学研究科 間瀬茂

家系図データに基づく連鎖解析は疾患原因遺伝子の座位を特定するための基本的手法である。家系図データはしばしばメンバーの遺伝子情報を欠き、また各座位の遺伝子対は一般にどちらが父・母親由来か直接決定できないという点で不完全である。関係遺伝子が全て共優性とし、順序情報を持つ (欠いた) 遺伝子対を遺伝子型 (表現型) と呼ぶことにし $[a, b]$ ((a, b)) と表す。連鎖解析の目標は複数座位間の遺伝子組替え率の推定であるが、そのためには各座位での家系図メンバーの遺伝子型同時分布や、家系内の親子間で遺伝子がどのように伝わったかを示す二値ベクトルである継承ベクトルの同時分布 (継承分布) を計算する必要がある。連鎖解析の代表的ソフトウェアである GENEHUNTER では、継承分布を家系図の創始者のあらゆる遺伝子型の組み合わせが特定の継承ベクトルを持つかどうかを総当たりで照合判定する堅実であるが、愚直な方法で行っており、このことがある程度大きな家系図の解析を困難にする原因となっている。この講演では BP (Belief Propagation, 信念伝播) や LBP(Loopy Belief Propagation) と呼ばれる新しいアルゴリズムを連鎖解析に用いる可能性について報告する。

BP は 1980 年代にアメリカの人工知能研究グループにより、決定論的エキスパートシステムに確率的依存関係を導入した BN(Bayesian Network、確率ネットワーク) に関する手法としてヒューリスティックに導入され、その後 Lauritzen 等の統計家により DAG(Directed Acyclic Graph、ループを持たない有向グラフ) の各ノードの周辺分布を高速計算するアルゴリズムとして整備された。一方で BP はループを持つグラフに於いてもしばしば有効であることが知られていたが、その場合このアルゴリズムが実際何を行っているかは 2001 年に Jedidia 等が統計力学のイジングモデル理論で古くから知られていたギブス自由エネルギーの Bethe 近似の極少化条件と BP の形式的同一性を指摘することにより、始めて明確な解釈を得た。ループのあるグラフ構造に対する BP とその変種を総称して LBP と呼ぶ。BP と同様に LBP は隣接ノード間でメッセージと呼ばれる量を交換することをグラフに沿って何度も繰り返す手法で、DAG の場合を除けばその収束性と収束値が各ノードの真の周辺分布を与えるかは必ずしも保証されない (ある LBP は適当な条件の下で収束性を保証する)。

ノード集合 $\{i\}$ を持つグラフ G の各頂点に有限離散値を取る確率変数 X_i が対応するとし、近傍関係 (辺でつながる) にあるノード対を $i \sim j$ と表す。$X = \{X_i\}$ の同時分布が次のようないわゆる二体相互作用を持つギブス (ボ

ルツマン) 分布で表現されるとする:

$$p(x_1, x_2, \ldots) = C \prod_{i \sim j} \psi_{ij}(x_i, x_j) \prod_i \phi_i(x_i \mid y_i)$$

ここで C は正規化定数 (一般に計算不可能) で、$\phi_i(x_i \mid y_i)$ はデータ y_i を観測した上での X_i の事後確率に比例する量である。ϕ_i, ψ_{ij} を統計物理学の用語を借りポテンシャル関数と呼ぶ。

LBP アルゴリズムはノード i からその近接ノード j へのメッセージと呼ばれる関数 $m_{ij}(x_j)$ を、初期値 1 から出発し次の更新式によりグラフの各辺 (i, j) に沿って一順することを収束するまで繰り返すことに相当する:

$$m_{ij}(x_j) = \sum_{x_i} \psi_{ij}(x_i, x_j) \phi_i(x_i \mid y_i) \prod_{\substack{k: k \sim j \\ k \neq i}} m_{kj}(x_j).$$

各ノード i および近接ノード対 i, j に於ける p の周辺分布は (比例定数を除いて) 次式で計算される:

$$b_i(x_i) = k \phi_i(x_i \mid y_i) \prod_{k: k \sim i} m_{kj}(x_j),$$

$$b_{ij}(x_i, x_j) = k \psi_{ij}(x_i, x_j) \phi(x_i \mid y_i) \phi(x_j \mid y_j) \prod_{\substack{k: k \sim i \\ k \neq j}} m_{ki}(x_i) \prod_{\substack{l: l \sim j \\ l \neq i}} m_{lj}(x_j)$$

b_i, b_{ij} からは必ずしも p を復元できないことを注意する。BP,LBP アルゴリズムの魅力はそれがグラフの辺に沿ったノード数に比例するローカルな算法からなっており、各ノードの値 x_i の組合せが計算量的に爆発する際にも適応可能なことにある。

家系図データでは両親とその一人の子供という三体関係が存在する。この場合に上記の LBP アルゴリズムを適応するためには、夫 i、妻 j に対し人工的に夫婦ノード (i, j) を導入し、夫 i、妻 j、その子供 k が (i, j) と近傍関係にあるとする。そしてポテンシャル関数は次のように決める:

1) 個人 i の一体ポテンシャルは表現型データ y_i の下での i の遺伝子型の事後確率、
2) 夫婦ノードの一体ポテンシャルは $\equiv 1$,
3) 夫、妻と対応する夫婦ノード間の二体ポテンシャルは $\equiv 1$,
4) 夫婦ノードと一人の子供間の二体ポテンシャルはメンデルの法則で決まる遺伝確率

この場合個人および夫婦の一体周辺分布と、親とその子供の二体周辺分布が求まれば、家系図全体の遺伝子型同時分布が復元可能であり、それより継承分布自身を計算することが可能である。

LBP アルゴリズムをループのある (つまり近親結婚を含む) 家系図データに適応した計算機実験の結果は講演中で紹介する予定である。

分子進化速度の確率変動モデルと適応進化の検出

岸野 洋久 (東大農), 梶谷康秀 (東大農)

1 ゲノム適応進化の諸相と時間スケール

生物の適応進化の背後にゲノムの変化がある。最も大きな変化はゲノムの持つレパートリーの多様化である。種々の遺伝子が重複を繰り返し、また稀にではあるが、ゲノムレベルで大規模に重複することもある (Ohno (1970); Gu (2002))。大部分は機能を失い偽遺伝子となるが、新たな機能を獲得する。また、遺伝子の部分機能化モデルは重複遺伝子が長期にわたり保持されている現象を良く説明している (Force *et al.* (1999))。さらに、バクテリアと藻類からなる原核生物においては、種を跨いだ遺伝子の水平伝播がしばしば観察されている。たとえば種々の抗生物質に対してこれまで例外なく耐性遺伝子が生まれ、水平伝播を通じて急速に広範囲の微生物に浸透して来た (Normark and Normark (2002))。

こうしたマクロ進化は大規模な変革をもたらすが、頻度はそう多くはないであろう。これに対して、点突然変異を主体とした遺伝子レベルの変化は、生物の絶えざる進化と多様化に大きな貢献をしてきたと考えられている。中でも、アミノ酸をコードするコード領域は、比較的保存性が高いこともあり、良く調べられている。適応進化を検出するための統計量として現在最もしばしば用いられているものが、非同義置換と同義置換の比である。機能的な制約からこの比は1より小さいことが多いが、免疫系に関与する遺伝子、病原菌への抵抗性遺伝子、性決定遺伝子などで、正の淘汰圧（多様化選択）が起きたことが認められている (Hughes and Nei (1988); Bishop *et al.* (2000); Swanson and Vacquier (1995))。同義置換は比較的頻繁に起こるため、比較的近縁の種の比較や種内変異の解析が適している。

さらに近年注目されているのが、遺伝子の発現量と発現パターンの調節による進化と多様化である。トウモロコシは、側枝の成長を抑制する遺伝子 *tb1* の調節領域に強い淘汰圧を受け、この遺伝子が過剰に発現した結果、その代償として人間にとって重要な穀物を齎した (Doebley *et al.* (1997))。ハワイのマウイ島に群生する silver sword は、北米の同種に比して顕著な多様性を示している。この植物の花弁形成を支配するパスウエイ上の遺伝子を調べたところ、構造遺伝子には淘汰圧は見られず、調節遺伝子に正の淘汰圧が見られた (Barrier *et al.* (2001))。遺伝子の発現に関連する変化は非常に速く、多くが種内変異を見ることにより検出される。特に栽培化された作物は、強い選択圧を受けて短い時間に変異を受けている可能性があるため、発現調節の変化を齎した要因を低雑音で検出できる可能性を持つ点で、関心を持たれている。粘り気のあるイネは、Waxy 遺伝子の第一イントロンの 5' スプライス部位が GT から TT に塩基置換することにより誕生した (Hirano *et al* (1998))。

2 分子系統樹の尤度と進化速度の確率変動モデル

ゲノムに残された生物の適応のプロセスを推測するためには、どの遺伝子のどの部分で、またどの系統で分子進化速度が大きく変化したか、感度良く検出することが重要である。進化速度が確率変動を事前分布として取り込んだ階層モデルは、適応に伴う速度の変化を高感度で検出する可能性を持っている。

進化は座位間で独立とすると、相同な s 本の配列の対数尤度は

$$l(\theta|\mathbf{X}) = \sum_{h=1}^{n} \log f(\mathbf{X}_h|\theta) \tag{1}$$

と表わされる。ここで $\mathbf{X}_h = (X_{1h}, \ldots, X_{sh})'$ は第 h 座位のデータ、$\boldsymbol{\theta}_i$ は進化のプロセスを規定するパラメータである。配列の変化は、マルコフ過程によりモデル化する。分岐後それぞれの種の配列は独立に進化すると仮定すると、$f(\mathbf{X}|\theta)$ は

$$f(\mathbf{X}|\theta) = \sum_{Z_{i_0}} \pi_{Z_{i_0}} \prod_{j \in node(T) \setminus i_0} \sum_{Z_j} P_{Z_{anc(j)}, Z_j}(t_{anc(j),j}) \tag{2}$$

と簡単に表される。ここで、$node(T)$ は系統樹 T の節を表し、i_0 はその根である。$anc(j)$ は j に隣接する祖先となる節である。$P_{xy}(t)$ は時間 t を経た推移確率である。推移速度行列を $\mathbf{R}$ と書くと、推移確率行列は $P(t) = \exp(t\mathbf{R})$ として求められる。

速さ 1 に規格化された速度行列 $\mathbf{R}_0$ は系統間で異ならず、$\mathbf{R} = r\mathbf{R}_0$ (r はスカラー) としたとき、r のみが確率変動するモデルを考える (Thorne *et al.* (1998); Kishino *et al.* (2001))。進化速度の事前分布として、$r(t)$ の対数をとった $\tilde{r}(t) = \log r(t)$ に対して、次の 1 次-2 次のモーメント

$$\begin{aligned} E\left[\tilde{r}(t)|\tilde{r}(s)\right] &= \tilde{r}(s) - \frac{\nu}{2}(t-s) \\ V\left[\tilde{r}(t)|\tilde{r}(s)\right] &= \nu(t-s) \qquad (t > s) \end{aligned} \tag{3}$$

を持つ正規マルコフ過程を導入する。分岐後、速度は 2 系統で独立に変化するとするが、ウイルスの進化の解析では系統間の相関も重要となる (Shankarappa *et al.* (1999))。分岐年代の事前分布はガンマ分布とディリクレ分布により表現する。複数の遺伝子を同時解析することにより、共進化を検出することも可能となった (Thorne and Kishino (2002))。

ところで遺伝子の得失は、ミクロ進化速度が加速した極限とみなすことができる。遺伝子の有無に関する 2 値マルコフ過程を考えることにより、このプロセスをモデル化することができる。原核生物 42 種のゲノムを比較し、3165 のたんぱく質の共有関係を調べたところ、情報関係やリボソームの構造など生命の機構に本質的な遺伝子は速度が遅く、代謝関係や転写の調節に関与する遺伝子は速いことが浮き彫りになった。また、16S rRNA との関係において速度の変化を解析したところ、バクテリアの共生とともに大規模に遺伝子が抜け落ちている様子などが明瞭に読み取れた。

参考文献

Barrier M, Robichaux RH and Purugganan MD (2001). Accelerated regulatory gene evolution in an adaptive radiation *Proc. Natl. Acad. Sci., USA.* **98**: 10208–10213.

Bishop JG, Dean AM and Mitchell-Olds T (2000). Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. *Proc. Natl. Acad. Sci. U. S. A.* **97**: 5322–5327.

Doebley J, Stec A and Hubbard L (1997). The evolution of apical dominance in maize. *Nature.* **386**: 485–488.

Force A *et al* (1999). Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics.* **151**: 1531–1545.

Gu X, Wang Y and Gu J (2002). Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nature Genetics.* **31**: 205–209.

Hirano HY, Eiguchi M and Sano Y (1998). A single base change altered the regulation of the Waxy gene at the posttranscriptional level during the domestication of rice. *Mol. Biol. and Evol.*, **15**: 978–987.

Hughes AL and Nei M (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature.* **335**: 167-170.

Kishino H, Thorne JL, and Bruno WJ (2001). Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. and Evol*, **18**: 352–361.

Normark BH, Normark S (2002). Evolution and spread of antibiotic resistance. *J. Intern. Med.* **252**: 91–106.

Ohno S (1970) *Evolution by gene Duplication.* Springer-Verlag, Heidelberg, Germany.

Shankarappa R *et al* (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of Virology.* **73**: 10489–10502.

Swanson WJ and Vacquier VD (1995). Extraordinary divergence and positive Darwinian selection in a fusagenic protein coating the acrosomal process of abalone spermatozoa. *Proc. Natl. Acad. Sci. U. S. A.* **92**: 4957-4961.

Thorne JL and Kishino H (2002). Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.* **51**: 689–702.

Thorne JL, Kishino H and Painter IS (1998). Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**: 1647–1657.

SNP同定の2段階試験デザインにおけるサンプルサイズ設計

佐藤泰憲　寒水孝司　吉村功
東京理科大学工学研究科

1　はじめに

ヒトゲノムの全塩基配列が2000年6月にほぼ解読されたことをきっかけに，ポストゲノム計画として多くの研究が盛んに行われてきている．

そのような研究の1つにSNP解析があるが，その目的は疾患に関連する遺伝子多型を同定することにある．SNP解析は効率よく安価な方法であることが望ましく，試験デザインに対してもいくつかの検討すべき課題が考えられる．

本研究では，疾患に関連する遺伝子多型を同定することを目的としたケースコントロール研究のサンプルサイズ設計に着目し，Satagopanら[1]の論文で提案されているコストを制約とした2段階試験デザインのサンプルサイズ設計を取り上げる．そこで，著者らが提案したサンプルサイズ設計の問題点を明らかにするとともに，2段階試験デザインのサンプルサイズ設計を拡張する．

2　2段階試験デザイン

2段階試験では，まずステージ1でn_1人からそれぞれm個のSNPを調べて，評価値が最大のものから$m\gamma$番目までのSNPをステージ1での有望なSNPとする．ここで，γはステージ2で測定するSNPの割合$(0<\gamma<1)$である．次にステージ2で，ステージ1とは異なるn_2人からそれぞれステージ1で有望とされたSNPを$m\gamma$個調べる．そしてステージ1とステージ2を合わせて評価値が最大のSNPを疾患と関連があるSNPと同定する．試験を2段階にする理由は，コストを制限した場合には，1段階の試験を行うよりも疾患と関連が高いSNPをより高い確率で同定できるからである．

コストは，人数とSNPの数の積に比例するので，総コストは$T=n_1m+n_2m\gamma$となる．(ここでの総コストは，比例定数を除いたものである.) ここで，ステージ1で用いるコストの割合を$\omega=n_1m/T$とする．そして，Tとmを固定し，2段階試験において疾患と最も関連があるSNPが同定される確率（以下，正判定率と呼ぶ）が最大になるγとωを求めることを考える．2段階試験の前提条件として，疾患と関連があるSNP（疾患関連SNP）は1個，疾患と関連がないSNP（非疾患関連SNP）は$m-1$個とする．

SNPごとにケースとコントロールの比率の差をその標準誤差で割り，正規近似した値を疾患とSNPとの関連を表す評価変数とする．SNPの評価変数は互いに独立とすると，評価変数に加法性が成立するので，ステージ1とステージ2を合わせた評価変数を基準として，疾患に最も関連するSNPを同定することができる．

3 研究課題

著者らは，ステージ1で全体のコストの75%（$\omega = 0.75$）を使ってすべてのSNPをスクリーニングし，ステージ2で残りのコストを使って10%（$\gamma = 0.10$）の有望なSNPを調べれば多くの場合について正判定率がほぼ最大になると結論付けた．ところが，著者らによるγとωの検討は正判定率が比較的高くなる状況に限られており，正判定率が低くなる状況について十分な検討がなされていない．そこで，著者らが検討した状況よりも広い範囲でも正判定率が最大になるかをシミュレーションにより検討した．また，疾患関連SNPが複数の場合を想定して，新たな正判定率を定義した．その正判定率の比較も検討した．

さらに，2段階試験の状況設定はやや単純すぎる側面がある．例えば，評価変数は比率の差を正規近似したもので，疾患とSNPの関連を正しく評価していない可能性がある．そこで，現実的な状況に対応できるように評価変数を疾患とSNPのオッズ比に変え，ステージ1でSNPを有望と判定する際に基準値を導入することを提案した．そして，その性能をシミュレーションにより評価した．

4 まとめ

著者らが提案した2段階試験を再検討し，設定すべきパラメータは状況に応じて使い分けた方がよいことが分かった．また，評価変数をオッズ比に変えて新しい試験デザインを提案した．

しかし，多因子病を想定したSNP解析では環境要因の影響があるので，交絡因子の影響を考慮するために評価変数を調整オッズ比に変える必要がある．また，遺伝子間の相関を考慮することも必要である．これらは今後の課題として取り組んでいく．

参考文献

[1] Jaya, M. S., David, A. V., 2002 "Two-Stage Designs for Gene - Disease Association Studies" Biometrics, 58, pp. 163 - 170.

[2] Feller, W., 1966 " An Introduction to Probability Theory and Its Applications, Volume2 " New York:Wiley

[3] Zelen, M., 1971 " The analysis of several 2 × 2 contingency. " , Biometrika, 58, 1, pp.129 - 137.

[4] W.James, 2002 " Sample size requirements for matched case-controlstudies of gene-environment interaction. " Ststistics in Medicine, 21, pp.35 - 50.

[5] Risch, N. J., 1996 " The Future of Genetic Studies of Complex Human Diseases. " Science, 273, pp. 1516 - 1517.

代表者連絡先　吉村功　東京理科大学工学部経営工学科　〒162-8601 東京都新宿区神楽坂1-3
Tel：03-5228-8350　Fax：03-3260-5770
E-mail：isao@ms.kagu.tus.ac.jp

Copula Models in Multivariate Survival Analysis

Hideatsu Tsukahara
Department of Economics, Seijo University
(tsukahar@seijo.ac.jp)

December 7, 2002

1. Introduction. When we have two failure times T_1 and T_2 and would like to study correlation between them, we need families of two-dimensional survivor functions for statistical modeling. Cox and Oakes [2] list a number of desirable properties for these families:

(i) The association between T_1 and T_2 is governed by a single parameter θ which has a simple physical interpretation.

(ii) The marginal survivor functions can be specified arbitrarily and, if desired, parameterized separately from θ.

(iii) Either negative or positive association should be permissible, and the special cases of independence and the Fréchet-Hoeffding bounds are achievable within the family.

(iv) Reasonably simple parametric and semiparametric procedures are available for estimating θ, even in the presence of censoring in either or both components.

They suggest the use of Clayton copula models (Clayton [1]). But in fact, many other copula models are available, and we study some semiparametric estimation problems ((iv) above) in general copula models with bivariate right-censored data.

2. Copula Models. Let T_1 and T_2 be two failure times with joint continuous survivor function $S(t_1,t_2) = \mathbf{P}(T_1 > t_1, T_2 > t_2)$, and C_1 and C_2 be two censoring times with joint survivor function $U(t_1,t_2) = \mathbf{P}(C_1 > t_1, C_2 > t_2)$. What we observe is $(X_1, X_2, \delta_1, \delta_2)$, where $X_i = T_i \wedge C_i$ and $\delta_i = \mathbf{1}_{\{T_i \le C_i\}}$ for $i = 1, 2$. We assume that (T_1, T_2) and (C_1, C_2) are independent.

A copula is a multivariate distribution function with all univariate marginals being $U(0,1)$. By Sklar's theorem (see Nelsen [5]), we can find a 2-dimensional copula C satisfying $S(t_1,t_2) = C\big(S_1(t_1), S_2(t_2)\big)$, where S_1 and S_2 are the marginal survivor functions. C is called the survival copula of (T_1, T_2). The most well-known copula in survival analysis is the Clayton family mentioned above:

$$C(u_1,u_2) = (u_1^{1-\theta} + u_2^{1-\theta} - 1)^{1/(1-\theta)} \vee 0, \quad \theta \in [-1,0) \cup (0,\infty)$$

Copulas join multivariate survivor functions to their one-dimensional marginals. This indicates that we can separately model univariate marginals and dependence structure represented by copulas.

3. Semiparametric Estimation. Now let $(T_{1k}, T_{2k}, C_{1k}, C_{2k})$, $k = 1, \dots, n$ be iid copies of (T_1, T_2, C_1, C_2), and put $X_{ik} = T_{ik} \wedge C_{ik}$ and $\delta_{ik} = \mathbf{1}_{\{T_{ik} \le C_{ik}\}}$ for $i = 1, 2$; $k = 1, \dots, n$. Suppose that the survival copula of (T_1, T_2) belongs to a parametric family $\{C_\theta \colon \theta \in \Theta \subset \mathbb{R}^m\}$. Our problem is then to estimate θ based on the obsevations $(X_{1k}, X_{2k}, \delta_{1k}, \delta_{2k})$, $k = 1, \dots, n$ with unknown marginals. We shall discuss two estimation methods: rank approximate M-estimation and minimum distance method.

(1) *Rank approximate M-estimator.* Assuming the density c_θ of C_θ exists, we can write down the full likelihood. We replace the marginal survivor functions in the resulting likelihood equation to get the following estimating equation:

$$\sum_{k=1}^{n} \phi\big(\mathbb{S}_{n1}(X_{1k}), \mathbb{S}_{n2}(X_{2k}), \delta_{1k}, \delta_{2k}, \theta\big) = 0, \tag{1}$$

where $\mathbb{S}_{ni}$ be the product-limit estimator of S_i for $i = 1, 2$, and ϕ is some score function generalizing the efficient score $\dot{l}_\theta(u, v, \delta_1, \delta_2)$ which equals the partial derivative of

$$\delta_1\delta_2 \log c_\theta(u,v) + \delta_1(1-\delta_2)\log\frac{\partial C_\theta(u,v)}{\partial u} + (1-\delta_1)\delta_2\log\frac{\partial C_\theta(u,v)}{\partial v} + (1-\delta_1)(1-\delta_2)\log C_\theta(u,v)$$

with respect to θ. We call any solution $\widehat{\theta}_n^{RAM}$ to (1) a *rank approximate M-estimator*. Shih and Louis [6] prove the asymptotic normality of $\widehat{\theta}_n^{RAM}$ with $\phi = \dot{l}_\theta$ under somewhat strong assumptions (continuity and especially boundedness of certain derivatives). It is possible to relax their assumptions to allow wider family of copulas.

(2) *Minimum distance estimator*. The method of minimum distance is known to be robust, so when a slight deviation from a given parametric family is anticipated, it may be an appropriate method to employ. Also, the minimum distance estimator is consistent under very general conditions, so it may serve as an initial estimator for more involved estimation procedures. Let the *MD functional* T on the space of copulas be defined by

$$T(D) = \arg\min_\theta \int_0^1\int_0^1 [C_\theta(u,v) - D(u,v)]^2\, dudv$$

(some truncation of the range of integration may be technically necessary). Let $\mathbb{S}_n$ be a nonparametric estimator of the bivariate survivor function S (e.g., Dabrowska's or Prentice-Cai's. See Gill et al. [4]). A version of empirical copula for right-censored data may then be defined by $\mathbb{C}_n(u,v) = \mathbb{S}_n(\mathbb{F}_{n1}^{-1}(1-u), \mathbb{F}_{n2}^{-1}(1-v))$ for $(u,v) \in (0,1)^2$. Finally let $\widehat{\theta}_n^{MD} = T(\mathbb{C}_n)$, which we call an *minimum distance estimator*. It is not difficult to show the continuity and differentiability of the MD functional T, from which, with the results in Dabrowska [3] and Gill et al. [4], one can derive the asymptotic normality of $\widehat{\theta}_n^{MD}$.

4. Extensions. It is (theoretically) easy to extend the estimation methods discussed above to the higher dimensional case though the notation would be rather complicated. It would be the next project to incorporate possible left-truncation or interval censoring. Wang and Ding [7] study the pseudo-likelihood estimation with bivariate current status data, but they allow only single monitoring time. The minimum distance method can be easily applied once we find tractable estimators of joint and marginal survivor functions to construct the empirical copula.

References

[1] Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika*, **65**, 141–151.

[2] Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*, Chapman and Hall, London.

[3] Dabrowska, D. M. (1996). Weak convergence of a product integral dependence measure, *Scand. J. Statist.*, **23**, 551–580.

[4] Gill, R. D., van der Laan, M. J. and Wellner, J. A. (1995). Inefficient estimators of the bivariate survival function for three models, *Ann. Inst. Henri Poincaré - Probab. Statist.*, **31**, 545–597.

[5] Nelsen, R. B. (1999). *An Introduction to Copulas*, Lecture Notes in Statistics, Vol. 139, Springer-Verlag, New York.

[6] Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data, *Biometrics*, **51**, 1384–1399.

[7] Wang, W. and Ding, A. A. (2000). On assessing the association for bivariate current status data, *Biometrika*, **87**, 879–893.

Model checking for transformation model

服部 聡 *

2002年12月07日

1 Transformation modelの推測理論

T を連続分布を持つ failure time、C を potential censoring time とする。Z を p-次元共変量ベクトルとする。Z の各成分の絶対値は1で押さえられるとしておく。C の生存時間分布は共変量 Z に依存しないとし、G と書く。$X = min(T, C)$、$\Delta = I(T \leq C)$ とする。観測されるのは (X, Δ, Z) であり、$\{(X_i, \Delta_i, Z_i)\}_{i=1}^n$ をその i.i.d copy とする。共変量 Z を所与としたときの生存時間関数を $S_Z(t)$ とする。Transformation model は、

$$S_Z(t) = g\{h(t) + Z^T\beta\}$$

により定義される。ただし、$h(t)$ は未知の単調増大関数であり、g は既知の単調減少関数である。Cox 比例ハザードモデル、比例オッズモデルは transfomation model の特別な場合である。β, $h(t)$ の推定法が Cheng, Wei and Ying(1995 Biometrika) により提案された。

2 Model checking procedure

生存時間解析における自然な残差は martingale 残差であるが、Transformation model に対しては、$\Lambda_{Z_i}(t) = -log(S_{Z_i}(t))$ という関係を用いて、

$$\hat{M}_i(t) = N_i(t) + \int_0^t Y_i(u) dlog\big(g\{\hat{h}(t+) + Z_i^T\hat{\beta}\}\big).$$

として自然に martingale 残差が定義される。ここで $\hat{\beta}, \hat{h}(t)$ は Cheng, Wei and Ying(1995 Biometrika) による一致推定量である。Martingale 残差はその値域が著しく非対称であり、martingale 残差のプロットからモデルの適合を判断するのは困難である。Lin, Wei and Ying(1993 Biometrika) は Cox 比例ハザードモデルに対して、累積し対称化した martingale 残差によるモデルの適合の方法を開発した。本研究では、同様の方法により transformation model の当てはまりを評価する方法を考える。モデルの適合を測る検定統計量として、以下のような多次元パラメータ確率過程（確率場）を定義する。

$$H(t, z) = \sum_{i=1}^n I(Z_i \leq z)\hat{M}_i(t).$$

モデルが正しいと仮定すると、Taylor 展開により $n^{\frac{1}{2}}H(t, z)$ は漸近的に

$$n^{-\frac{1}{2}}\tilde{H}(t, z) \quad = \quad n^{-\frac{1}{2}}\sum_{l=1}^n \int_0^t I(Z_l \leq z) dM_l(u)$$

*北里大学大学院臨床統計、中外製薬株式会社臨床解析部

$$
\begin{aligned}
&+ \quad n^{-1}\sum_{l=1}^{n}\int_0^t I(Z_l \le z)Y_l(u) \\
&\times \quad d\Big[\tilde{g}\{h_0(u)+Z_l^T\beta_0\}W_{Z_l}(t)\Big].
\end{aligned}
$$

と同等となる。ここで、$\tilde{g}(x)=\frac{dg(x)}{dx}/g(x)$ であり、$W_{Z_l}(t)$ は $n^{\frac{1}{2}}\{\hat{h}(t)-h_0(t)+(\hat{\beta}-\beta_0)^T Z_l\}$ と漸近的に同等な zero-mean Gaussian process である（Cheng, Wei and Ying 1997 JASA）。このことを用いると、帰無仮説（モデルが正しい）下で、$n^{-\frac{1}{2}}\tilde{H}(t,z)$ は zero-mean Gaussian process に弱収束することを示すことができる。モデルの適合を評価するには、$n^{-\frac{1}{2}}H(t,z)$ の帰無分布を評価する必要があるが、この Gaussian process の相関関数を解析的に扱うのは困難である。ここで新たな確率過程

$$
\begin{aligned}
n^{-\frac{1}{2}}\hat{H}(t,z) &= n^{-\frac{1}{2}}\sum_{l=1}^{n}\int_0^t I(Z_l \le z)d\hat{M}_l(u)\mathcal{L}_l \\
&+ \quad n^{-1}\sum_{l=1}^{n}\int_0^t I(Z_l \le z)Y_l(u) \\
&\times \quad d\Big[\tilde{g}\{\hat{h}(u)+Z_l^T\hat{\beta}\}\hat{W}_{Z_l}(t)\Big]\mathcal{L}_l.
\end{aligned}
$$

を定義する。ここで、$\{\mathcal{L}_i\}_{i=1}^n$ はデータ $\{(X_i,\Delta_i,Z_i)\}_{i=1}^n$ と独立な標準正規確率変数とする。また、$\hat{W}_{Z_l}(t)$ は $W_{Z_l}(t)$ 中の理論値を一致推定量で置き換えたものとする。データ $\{(X,\Delta,Z)\}_{i=1}^n$ を固定すると、確率変数は $\{\mathcal{L}_i\}_{i=1}^n$ のみであり、$n^{-\frac{1}{2}}\hat{H}(t,z)$ は zero-mean Gaussian process となるが、データ $\{(X_i,\Delta_i,Z_i)\}_{i=1}^n$ を固定した下での条件付き分布が、$n^{-\frac{1}{2}}\tilde{H}(t,z)$ の条件付きでない分布に漸近的に同等であることが示せる。データを固定したとき、$n^{-\frac{1}{2}}\hat{H}(t,z)$ のサンプルパスは $\{\mathcal{L}_i\}_{i=1}^n$ を計算機内で発生させることで、容易にシミュレートできる。したがって、検定統計量 $n^{-\frac{1}{2}}H(t,z)$ の帰無分布は、シミュレーションにより近似することができる。$n^{-\frac{1}{2}}H(t,z)$ の帰無分布（プロセス）から任意の数の実現値を発生させることで帰無分布が評価でき、したがって $\sup_{z\in[-1,1]^p,0\le t\le\tau}|H(t,z)|$ というような統計量の p-値の評価も可能である。この統計量の 1 次元の sub-class で、共変量の関数形あるいは link 関数形の misspecification に対して志向性を有すると考えられる統計量を定義することも可能である。これらは 1 次元の確率過程であり、同様の simulation による方法により、graphical な model の評価が可能である。

Variance Components Testing in Mixed Effects Models

Fumiaki Takahashi
Division of Biostatistics

1 Introduction

Nonlinear mixed effects model has been applied for pharmakokinetics (PK) and pharmacodynamics (PD) as a standard statistical analysis. Population average (PA) approximation, which can be derived from the first-order Taylor approximation of the objective nonlinear function with respect to random effects evaluated at their expected values zero, have been used without checking assumption that an inter-subject variation is sufficiently small. We can reduce the problem as a one-side hypothesis test $H : \theta = 0$ against $K : \theta \geq 0$, where θ is a parameter vector which fully characterize variance components of the intra-random effects. Likelihood ratio method requires a complex numerical integration to estimate parameters of interest under the alternative hypothesis. In addition, the maximum likelihood approach faces mathematical difficulties such that the null hypothesis places the true value of the variance parameters on the boundary of the parameter space defined by the alternative hypothesis. This fact causes the limiting distribution of -2 times of the logarithm of the likelihood ratio as a mixture of χ^2 random variables instead of a regular χ^2 random variable. Score method is an alternative approach. The main advantages of the method are that 1) it requires only estimation of the model under the null hypothesis, 2) it is applicable even if the exact population distribution is unknown and 3) it is asymptotically equivalent to the likelihood ratio test statistic. Therefore the score method has a computational advantage in nonlinear mixed effects models. In this article we construct the score statistic testing for nozero variance components under the mild distribution assumptions in nonlinear mixed effects models. Via Monte Carlo simulation study we investigate effects of the size and the power of the score test with respect to those of the likelihood ratio test in various small sample and under the variation of distribution assumptions.

2 Score statistic

We define the nonlinear mixed effects model as follows: $y_{ij} = f(x_{ij}, \beta, b_i) + \epsilon_{ij}$ for $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$, where y_{ij} is a jth observation on ith subject, $f(\cdot)$ is a nonlinear function with existence of third derivatives, x_{ij} is a known design matrix, β is an unknown fixed effect parameter vector$(P \times 1)$, b_i is an unobserved inter-random effect vector$(Q \times 1)$, and ϵ_{ij} is a intra-random error. We assume that y_i are mutually independent given that b_i, and ϵ_i and b_i are mutually independent. Further we assume that an inter-random effect vector b_i is a random sample from a distribution function F with mean zero and covariance matrix $D(\theta)$, where θ is a parameter vector$(M \times 1)$ which fully characterize $D(\theta)$. We assume that an intra-random error vector ϵ_i is a random sample from a population with mean zero and covariance matrix R_i. Then the marginal quasi-likelihood for all subjects, L is given by $L(\beta, \theta) \propto \prod_{i=1}^{K} \int \exp\left[l_i\left(X_i; \beta, b_i\right)\right] dF(b_i; \theta)$,where $l_i(X_i; \beta, b_i) \propto \int_{y_i}^{\mu_i^{b_i}} R_i^{-1}(y_i - u)\, du$ and $\mu_i^{b_i} = E(y_i) = f(X_i; \beta, b_i)$. To avoid this numerical integration to obtain the marginal quasi-likelihood, we consider that the integrand can be approximated in a neighborhood of the null hypoth-

esis $H_0 : \theta = 0$, which is equivalent to $H_0 : b_i = 0$. Evaluating the integration by taking the expectation with respect to b_i, we have the marginal log-quasi-likelihood for all subjects; $l(\beta, \theta) \propto \sum_{i=1}^{K} \left\{ l_i^0 + \frac{1}{2} tr V_i D(\theta) \right\} + o(\|\theta\|)$ where $l_i^0 = l_i(X_i; \beta, b_i = 0)$ and V_i is a sum of an outer product of the first derivative of $f(X_i; \beta, b_i)$ with respect to b_i and the second derivative of $f(X_i; \beta, b_i)$ with respect to b_i evaluated at the null hypothesis $b_i = 0$. To check for testing $H_0 : \theta = 0$ we construct a score statistic $\chi_G^2 = U_\theta(\hat{\beta})^T \tilde{I}(\hat{\beta})^{-1} U_\theta(\hat{\beta})$ where $U_\theta\left(\hat{\beta}\right) = \partial l(\beta, \theta) / \partial\theta|_{\beta=\hat{\beta}, \theta=0}$, $U_\beta\left(\hat{\beta}\right) = \partial l(\beta, \theta) / \partial\theta|_{\beta=\hat{\beta}, \theta=0}$, $I_{\theta\theta} = E\left[U_\theta\left(\hat{\beta}\right) U_\theta^T\left(\hat{\beta}\right)\right]$, $I_{\theta\beta} = E\left[U_\theta\left(\hat{\beta}\right) U_\beta^T\left(\hat{\beta}\right)\right]$, $I_{\beta\beta} = E\left[U_\beta\left(\hat{\beta}\right) U_\beta^T\left(\hat{\beta}\right)\right]$, and $\tilde{I}(\hat{\beta}) = I_{\theta\theta} - I_{\theta\beta}^T I_{\beta\beta}^{-1} I_{\beta\theta}$. Here $\hat{\beta}$, the maximum likelihood estimator of β under $\theta = 0$, can be easily obtained by fitting the nonlinear model. It can be shown that under the mild regularity conditions, the score statistic χ_G^2 asymptotically follows a chi-squared distribution with M degree of freedom under the null hypothesis; $\theta = 0$.

3 Results

Monte Calro simulation studies in a simple linear and nonlinear mixed effects model show the following results: 1)The size of score statistics remarkably depend on the number of observations per subject, and is about 0.030 to 0.045 when the number of observations per subject is five or more. Note that the number of subjects does not affect the size compared with the number of observation per subject. 2)The size of score statistic is closer than that of likelihood ratio statistic based on a χ^2 random variable to a nominal size when the number of observations per subject is five or more. The size of likelihood ratio statistics based on a mixture of χ^2 random variables is constantly closest to a nominal size. However it frequently exceeds a nominal size. 3)The power of score statistic is superior to that of likelihood ratio statistic base on a χ^2 random variable near the null hypothesis. The likelihood ratio statistic based on a mixture of χ^2 random variables is most powerful. 4)Score statistic is robust for the variation of a distribution assumption compared with likelihood ratio statistic.

4 Conclusion and Discussion

We investigate the statistical properties of the derived score test for nonzero variance components with mild assumptions of the existence of the first two moments about each of observations and random effects. Unfortunately the size and power of the score test is inferior to those of likelihood ratio test based on a mixture of χ^2 random variables via the simulation studies. However, especially in a nonlinear setting, the score test approach has a large computational advantage. We do not need calculate a likelihood under the alternative hypothesis and a weight of mixture of χ^2 random variables. We can apply this method to many situations. For example, recently worldwide simultaneous drug development has been conducted briskly by the impact of ICH. When normal assumption of a primary endpoint is invalid with various estimates of drug effect among countries considered as random effects, we can apply generalized linear mixed effects model to the data. Similarity of the derived effect among each country becomes very important issue for new drug applications (NDA) to regulatory agencies. It can be shown by the test of the null hypothesis through the proposed score test approach. We think that we need further researches on an asymptotic distribution of the score statistic and a method of estimating nuisance parameters to improve the size and power via actual applications.

2002年12月5日-7日
駒場シンポジウム「バイオスタティスティックスの数理的基礎」

cDNAマイクロアレイと実験計画法

濱野鉄太郎（北里大学大学院　薬学研究科　臨床統計部門）

概要

cDNAマイクロアレイは，臨床検体あるいは細胞株の遺伝子発現情報を，一度に大量に定量化する道具である．cDNAマイクロアレイ実験は高精度な技術に基づくものであり，微細な実験変動が結果に大きく影響する．そのため，cDNAマイクロアレイ実験の精度と再現性を向上させるための試みが多く行われており，実験計画法の適用もそのひとつである．本発表では，cDNAマイクロアレイ実験を実験計画法の観点から吟味する．

cDNAマイクロアレイ

近年，医学および生物学の分野において遺伝子発現解析が数多く行われている．その目的は，生命現象および疾患を遺伝子発現の観点から理解すること，およびその結果を社会に応用することである．例えば癌細胞の遺伝子発現解析は現在最も精力的に行われている研究のひとつであり，細胞が癌化に至る仕組の解明や分子標的薬の開発に有益な情報を提供している．このような遺伝子発現解析が幅広く行われるようになった理由のひとつに，科学技術の進歩に伴い，一度に大量の遺伝子発現を定量できる道具が開発されたことがある．そして，そのような道具のひとつがcDNAマイクロアレイ [1] である．

cDNAマイクロアレイは，スライドガラス上に何千から何万の相補的DNA(cDNA)を配置したものである．サンプルから抽出したメッセンジャーRNA(mRNA)をcDNAに逆転写し，cDNAマイクロアレイとハイブリダイズすることにより，サンプル内の横断的な遺伝子発現情報を一度に得ることができる．たとえば，ある種の癌細胞の遺伝子発現プロファイルをcDNAマイクロアレイで定量し，正常細胞のそれと比較することにより，ある種のがん細胞で特異的に発現する遺伝子群を特定することができる．

cDNAマイクロアレイでは，各スポットに配置されたcDNA量にばらつきがあるため，競合的ハイブリダイゼーションを行うのが一般的である．競合的ハイブリダイゼーションとは，2種類の異なるサンプルをそれぞれ異なる蛍光色素で標識化して，それらを混合したものをcDNAマイクロアレイ上でハイブリダイズさせることである．競合的ハイブリダイゼーションを行い，2種類のサンプル間の相対的な遺伝子発現比を定量することで，各スポットのcDNA量のばらつきの影響を除去することができる．

cDNAマイクロアレイを用いた遺伝子発現解析では，解析対象のサンプルを一方の蛍光色素で標識化し，（解析対象ではない）対照のサンプルをもう一方の蛍光色素で標識化して，解析対象のサンプルと同数のマイクロアレイを用いて競合的ハイブリダイゼーションを行う実験計画が一般的である．たとえば，癌細胞と正常細胞の遺伝子発現を比較したいとき，従来の実験計画では，複数の癌細胞および複数の正常細胞を一方の蛍光色素で標識化し，それとは別に，対照のサンプルをもう一方の蛍光色素で標識化して，競合的ハイブリダイゼーションを行うことが多い．

実験計画法

実験計画法は1920年代からR.A. Fisherらによって創始された統計的方法である．Fisherは，ロザムステッド農業試験場の技師であった頃に，実験において誤差を制御するために行わなければならない原則として，

1. 反復（replication）
2. ランダム化（randomization）
3. 局所管理（local control）

を提唱した．そしてこの原則を満たすように実験を配置する方法論を展開した．

cDNA マイクロアレイと実験計画法

近年，その実験計画法の観点から cDNA マイクロアレイを見直す試みが行われている [2]．cDNA マイクロアレイはミクロレベルのアレイ製造技術に基づいて作成されていることから，微細な変動要因が実験結果に大きく影響する．アレイの製造からデータの取り込みまでの過程で細心の注意が必要なのはいうまでもないが，再現性のある結果を得るためには，実験計画法の観点からも精度の向上を図ることが重要である．ここでは，上記の Fisher の 3 原則に基づいて，cDNA マイクロアレイ実験を吟味する．

反復

cDNA マイクロアレイ実験で反復データを得るための方法は幾つか存在する．そのうち代表的なものとして，[3] で挙げられているものに加え，以下の 5 つが考えられる．

1. スポットの反復
2. 蛍光色素の反復（アレイおよび mRNA の調製が同じ）
3. マイクロアレイの反復（mRNA の調製が同じ）
4. マイクロアレイの反復（同一の検体あるいは細胞株だが，mRNA の調製が異なる）
5. マイクロアレイの反復（検体あるいは細胞株が異なる）

1 により，スポット内の cDNA 量のばらつきから生じる変動を低減できることに加え，反復スポットをなるべく遠くに配置することで，スポットのアレイ上での位置によって生じる系統的な変動を吟味することが可能である．2 によって，蛍光色素間での取り込み効率の違いを評価することができる．3 によって，ハイブリダイゼーション時に生じる実験的変動を低減することができ，またアレイ間でのデータの再現性を吟味することができる．4 によって，mRNA の調製を含む実験的変動を低減することができる．そして 5 は，実験結果を一般化する際に必要である．これらの反復を実験の目的に応じて使い分けることにより，結果の精度を向上させ，再現性を吟味することが可能になる．

ランダム化

前述したように，cDNA マイクロアレイ実験では微細な変動が結果に大きく影響する．たとえばマイクロアレイデータに日間変動が生じることが指摘されている [3]．このため，あるサンプルについては実験期間の前の方で解析し，他のサンプルを後の方で解析すると，サンプルと日間差が交絡してしまい，誤った結論を導く可能性がある．そのような系統的変動に対処するためには，実験の順番をランダムにすることが効果的である．しかし，マイクロアレイ実験においてランダム化を行うことの重要性はあまり認識されていない．ほかに，マイクロアレイ上に配置される遺伝子群は，機能の類似しているものが並列されている傾向があるが，これがスポットのアレイ上の位置によって生じる差と交絡する可能性がある．アレイ上のスポットの配置もできる限りランダムにすることで，そのような変動を制御することが可能である．

局所管理

cDNA マイクロアレイ実験では，解析対象のサンプルと（解析対象ではない）対照のサンプルで競合的ハイブリダイゼーションを行う実験計画が主流であるが，比較すべきものを対にする局所管理の観点からは，Kerr らが提案したループ計画の方がより好ましいであろう [2]．また，発現プロファイルを比較すべき遺伝子群が存在する際は，それらをアレイ上の近接した位置に配置することも考えられるが，上記のランダム化とバランスをとって配置する方法論が必要であろう．

参考文献

1 Eisen, M., Spellman, P., Brown, P. & Botstein, D. (1998) Proc. Natl. Acad. Sci. USA, 95, 14863 - 4868.

2 Kerr, MK. & Churchill GA. (2001) Biostatistics, 2 163-201.

3 Janssen, K. eds., (2003) *DNA Microarrays*, Cold Spring Harbor Laboratory Press, New York.

分子標的薬剤第Ⅰ相試験における遺伝子発現の用量反応性の解析

伊藤 陽一　大橋 靖雄

東京大学大学院医学系研究科生物統計学

1. 分子標的薬剤における第Ⅰ相試験

分子標的薬剤とは、癌細胞の増殖とその進展(運動、浸潤、転移)に関わる分子を標的とした薬剤である[1]。本発表で用いるデータはある分子標的薬剤の第Ⅰ相試験のデータである。通常抗癌剤の第Ⅰ相試験では、最大耐用量および薬物動態が調べられるが、本試験ではこれに追加し、薬剤が標的となっている遺伝子群およびその周辺遺伝子の発現量を変化させ得るのかということが調べられた。これは実験動物と人体では薬剤に対する遺伝子発現の反応プロファイルが往々にして異なるためであり、人体内における遺伝子発現プロファイルを調べることが、分子標的薬剤の抗腫瘍効果の機序の解明と本分子標的薬剤が奏効すると予想されるサブグループの特定のために必須となるためである。本試験で用いられたマクロアレイは機能が既知の遺伝子のみ775個スポットしたものであり、各遺伝子は機能別に12のグループに分けられていれる。各投与量ごとの患者数は100mgが3名、200mgが3名、300mgが6名、400mgが3名であった。このうち100mgの1名の試料の状態が悪く使えなかったため、全体のサンプルサイズは14名である。測定時点は投与開始前、投与開始後2日目、投与開始後8日目の3時点である。したがって、各遺伝子の投与開始後と投与開始前の比について、用量反応性があるかどうか探索することが本解析の目的である。

2. Biplotと用量反応性スコア

薬剤投与量と遺伝子発現との関連性を検討する探索的方法として、主成分分析の応用であるBiplot[(2)]を適用する。また、Biplotから導かれる用量反応性スコアを提案する。Biplotでは、対象者n人、遺伝子数p個のn×pデータ行列$\mathbf{X}$に関して以下のような階数分解[(3)]を考える。

$$\mathbf{X} = \mathbf{G}\mathbf{H}'$$

$\mathbf{G}$はn行r列、$\mathbf{H}$はp行r列の行列である。従って行列$\mathbf{X}$のk行l列の要素x_{kl}は$x_{kl} = \mathbf{g}_k'\mathbf{h}_l$と表わすことができる。$\mathbf{g}_k'$は行列$\mathbf{G}$の第$k$行ベクトルであり、$\mathbf{h}_l$は行列$\mathbf{H}$の第$l$行ベクトルである。ベクトル$\mathbf{g}_k'$や$\mathbf{h}_l$はそれぞれ第$k$対象者や第$l$遺伝子の発現に共通するベクトルであるため、第$k$対象者や第$l$遺伝子の発現に関する情報を要約したものと考えられる。しかし、上記の階数分解は一意ではないため、以下のような特異値分解を考える

$$\mathbf{X} = \mathbf{U}\ddot{\mathbf{A}}\mathbf{V}' = \sum_{i=1}^{r} \delta_i \mathbf{u}_i \mathbf{v}_i'$$

ここで$\ddot{\mathbf{A}}$は$\delta_1 \geq \delta_2 \geq \Lambda \geq \delta_r > 0$となる特異値$\delta_i$を対角要素として持つ$r$行$r$列の対角行列である。また、行列$\mathbf{U}$は列ベクトル$\mathbf{u}_1, \mathrm{K}, \mathbf{u}_r$からなる$n$行$r$列の行列で、$\mathbf{UU}' = \mathbf{I}$である。また、行列$\mathbf{V}$は列ベクトル$\mathbf{v}_1, \mathrm{K}, \mathbf{v}_r$からなる$p$行$r$列の行列で、$\mathbf{V}'\mathbf{V} = \mathbf{I}$である。ここで$\mathbf{G} = \mathbf{U}$，$\mathbf{H}' = \ddot{\mathbf{A}}\mathbf{V}'$とおく事によって$\mathbf{G}$と$\mathbf{H}$を一意に決定することができる。また以下のように、$\delta_2$までの和を取ることによって、もとのデータ行列$\mathbf{X}$を2次元空間内のベクトルの内積として近似することが可能となる。

$$\mathbf{X} \approx \delta_1 \mathbf{u}_1 \mathbf{v}_1' + \delta_2 \mathbf{u}_2 \mathbf{v}_2'$$

ここで第k対象者を(x, y)平面上に$\mathbf{u}_1, \mathbf{u}_2$の第$k$要素$u_{k1}, u_{k2}$を用いて、$(u_{k1}, u_{k2})$として布置する。この平面内において対象者が薬剤投与量の順に並ぶような方向を見つけることができれば、用量反応性に関する情報が得られるはずである。この方向を発見する探索的な方法として、個々の対象者の投与量を基準化し以下のような回帰式が考える。

$$(\text{基準化した投与量}) = \mathbf{u}_1 \beta_1 + \mathbf{u}_2 \beta_2 + \mathring{\mathbf{a}}$$

推定された回帰係数から単位ベクトル$\hat{\mathbf{a}} = (\hat{\beta}_1, \hat{\beta}_2) / \|(\hat{\beta}_1, \hat{\beta}_2)\|$を作り、用量反応性に関連する方向とする。この$\hat{\mathbf{a}}$と遺伝子ベクトル$\mathbf{h}_l$の内積は第$l$遺伝子の発現量のうち用量反応性に関連する部分を推定していると考えられる。これを用量反応性スコアと呼ぶことにする。この用量反応性スコアは遺伝子の生物学的解釈を行う前のスクリーニングに利用することができる。つまり、用量反応性スコアの大きい遺伝子から解釈を加えていけば良い。

この用量反応性スコアの問題点としては、一部の対象者の反応に強く依存してスコアが高くなっている可能性がある点が挙げられる。しかし、この問題に対しては用量反応性スコアに関してジャックナイフ推定[4]を行い、ジャックナイフ標本の変動係数と尖度がある一定値以上の遺伝子については解釈しないという立場を取ることによってある程度改善されると思われる。

参考文献

[1] 曽根三郎 分子標的薬剤とその臨床評価 オーバービュー. 現代医療. 2000: **32**; 2462-2470.

[2] Gabriel, K. R. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*. 1971: **58**; 453-467.

[3] Rao, C. R. Linear statistical inference and its application 2nd ed. John Whiley & Sons. 1973.

[4] Efron, B. and Tibshirani, R. J. An introduction to the bootstrap. Chapman & Hall. 1993.

遺伝子発現データに基づく遺伝子ネットワークの推定

井元 清哉[1]

Keywords: genetic network, microarray, Bayesian network, nonparametric regression, Bayes approach.

1 Introduction.

The microarray technology provides us enormous amount of valuable gene expression data. The analysis of the relationship among genes has drawn remarkable attention in the field of molecular biology and Bioinformatics. However, due to the causes of dimensionality and complexity of the data, it will be no easy task to find structures, which are buried in noise. To extract the effective information from microarray gene expression data, thus, new theory and methodology are expected to be developed from a statistical point of view. Our purpose is to establish a new method for extracting the relationships among genes clearer.

We consider constructing genetic network by using Bayesian network ([3]). To capture not only linear dependencies but also nonlinear structures between genes, we use nonparametric regression models with Gaussian noise. It is necessary to evaluate the estimated network by a suitable criterion. We derive a new criterion from Bayes approach ([1]). By using proposed method, we will overcome the defects of previous methods and attain more effective information. The efficiency of the proposed method is shown by the Monte Carlo simulation method. We also demonstrate our proposed method through the analysis of the *S. cerevisiae* cell cycle data [6].

2 Nonlinear Bayesian Network.

We consider a directed acyclic graph G and Markov assumption between nodes. Suppose that x_{ij} is a observation of i-th array of j-th gene and $p_{ik}^{(j)}$, for $k = 1, ..., q_j$ are observations of the parent genes of j-th gene. The joint density function is then decomposed into the conditional density of each variable, that is $f(\boldsymbol{x}_i) = \prod_{j=1}^{p} f_j(x_{ij}|\boldsymbol{p}_{ij})$, where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ and $\boldsymbol{p}_{ij} = (p_{i1}^{(j)}, ..., p_{iq_j}^{(j)})^T$. Then all we need to do is to consider how to construct the conditional densities $f_j(x_{ij}|\boldsymbol{p}_{ij})$.

In this paper, we capture the relationship between x_{ij} and $\boldsymbol{p}_{ij}$ by the nonparametric regression model $x_{ij} = \sum_{k=1}^{q_j} m_k(p_{ik}^{(j)}) + \varepsilon_{ij}$, $\varepsilon_{ij} \sim N(0, \sigma_j^2)$ and define the nonlinear Bayesian network model in the form

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \prod_{j=1}^{p} \exp[-\{x_{ij} - \sum_{k=1}^{q_j} \sum_{m=1}^{M_{jk}} \gamma_{mk}^{(j)} b_{mk}^{(j)}(p_{ik}^{(j)})\}^2/2\sigma_j^2]/\sqrt{2\pi\sigma_j^2},$$

where $m_k(p_{ik}^{(j)}) = \sum_{m=1}^{M_{jk}} \gamma_{mk}^{(j)} b_{mk}^{(j)}(p_{ik}^{(j)})$, $\boldsymbol{\theta}$ is a parameter vector, $b_{mk}^{(j)}(\cdot)$ are B-splines [2] of degree 3 and $\gamma_{mk}^{(j)}$ are coefficient parameters. If a gene has no parent genes, we substitute the B-spline structure by the constant function. It is clear that this model contains the linear regression model.

[1]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan. E-mail: {imoto, takao, miyano}@ims.u-tokyo.ac.jp

3 Criterion for choosing graph.

Let $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$ be the prior distribution on the unknown parameter vector $\boldsymbol{\theta}$ with hyper parameter vector $\boldsymbol{\lambda}$ and let $\log \pi(\boldsymbol{\theta}|\boldsymbol{\lambda}) = O(n)$. We construct the graph selection criterion based on the posterior probability of the graph. By using the Laplace approximation for integrals, we define the Bayesian network and nonparametric regression criterion, named BNRC, from Bayes approach

$$\begin{aligned} \mathrm{BNRC}(G) &= -2\log\left\{\pi_G \int \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})d\boldsymbol{\theta}_G\right\} \\ &\approx -2\log\pi_G - r\log(2\pi/n) + \log|J_\lambda(\hat{\boldsymbol{\theta}})| - 2nl_\lambda(\hat{\boldsymbol{\theta}}|\boldsymbol{X}_n), \end{aligned} \tag{1}$$

where π_G is the prior probability of the graph G, r is the dimension of $\boldsymbol{\theta}$, $\boldsymbol{X}_n$ is the microarray gene expression profile matrix,

$$l_\lambda(\boldsymbol{\theta}|\boldsymbol{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\log f(x_i;\boldsymbol{\theta}) + \frac{1}{n}\log\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}), \quad J_\lambda(\boldsymbol{\theta}) = -\partial^2 l_\lambda(\boldsymbol{\theta}|\boldsymbol{X}_n)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$$

and $\hat{\boldsymbol{\theta}}$ is the mode of $l_\lambda(\boldsymbol{\theta}|\boldsymbol{X}_n)$. The strategy of computing the mode $\hat{\boldsymbol{\theta}}$ is in [5]. The BNRC (1) depends on the hyper parameter $\boldsymbol{\lambda}$ and we optimize it based on BNRC. Hence, the optimal graph is automatically chosen such that the criterion BNRC (1) is minimal.

4 Computational experiments.

We examine the effectiveness of our method through the Monte Carlo simulation method. We set an artificial graph and the relationships between genes. The results of the Monte Carlo simulations show that our method can build the estimated graph, which is extremely close to the true graph.

We analyze the *S. cerevisiae* cell cycle data collected by [6]. This data was also analyzed by [4]. The results are that our method can represent many causal relationships, which agree with the knowledge of biology and the results of [4]. We could find some nonlinear dependencies, which linear models hardly find. The details of the Monte Carlo simulations and the analysis of the *S. cerevisiae* cell cycle data are shown in [5].

参考文献

[1] Berger J. O. 1985. *Statistical Decision Theory and Bayesian Analysis.* Springer, New York.

[2] de Boor C. 1978. *A Practical Guide to Splines.* Springer, Berlin.

[3] Cowell R., Dawid A., Lauritzen S. and Spiegelhalter D. 1999. *Probabilistic Networks and Expert Systems.* Springer, New York.

[4] Friedman N., Linial M., Nachman I. and Pe'er D. 2000. Using Bayesian Networks to Analyze Expression Data. *J. Comp. Biol.*, **7** 601-620.

[5] Imoto S., Goto T. and Miyano S. 2002. Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. *Proc Pacific Symposium on Biocomputing*, **7** 175-186.

[6] Spellman P., Sherlock G., Zhang M., Iyer V., Anders K., Eisen M., Brown P., Botstein D. and Futcher B. 1998. Comprehensive Identification of Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Mol. Biol. Cell*, **9**, 3273-3297.