Model Averaging without Non-negative constraints

奥井亮 & Guido Kuersteiner
京都大学 & Georgetown University

平成22年11月6日
IBIS2010

概要

- 線形回帰モデルにモデル平均手法を応用。
- 予測の平均二乗誤差を最小化する形でモデル平均をするための重み付けベクトルを選ぶ。
- 新しい点：　モデル平均をする重み付けベクトルの要素が負になってもよい。
- 利点１：　変数の順序にそれほどの制約がいらない。
- 利点２：　最適な重み付けベクトルの明示的な式をだすことができる。

設定

$\{y_i, x_i\}$: i.i.d. 標本。y_i: １変数。 x_i: 回帰変数のベクトルであり、可算無限個の変数からなる。 $x_i = (x_{1i}, x_{2i}, \dots)$.

無限個の回帰変数を含む線形回帰モデルを考える。なお分散均一性を仮定する。

$$
\begin{aligned}
y_i &= \mu_i + e_i, \\
\mu_i &= \sum_{j=1}^{\infty} \theta_j x_{ji}, \\
E(e_i|x_i) &= 0, \\
E(e_i^2|x_i) &= \sigma^2.
\end{aligned}
$$

例

この設定はどのような経済学的な状況を表現していると考えることができるか。

- ＧＤＰや株価の予測を多くの経済変数から行う。
- 経済成長率を多くの変数から説明する場合。
- シーヴを使ったノンパラメトリック回帰。

近似モデル

さて、真のモデルは、無限個の回帰変数が入っているのでこれを推定することはできない。そこで、ある有限個の変数のみを含むモデルを使って、真のモデルを近似し、yの予測をすることを考える。当然、近似モデルは一つではなく、いくつも考えることができる。

ここでは、$0 \leq k_1 < k_2 < \cdots < k_m < \ldots$を整数の数列として、$m$番目の近似モデルは最初の$k_m$個の回帰変数からなっている場合を考える。

各近似モデルは

$$y_i = \sum_{j=1}^{k_m} \theta_j x_{ji} + \eta_{mi} + e_i,$$

と書け、$\eta_{mi} = \sum_{j=k_m+1}^{\infty} \theta_j x_{ji}$が近似誤差である。

近似モデルの推定

最小二乗法を使って、近似モデルの推定を行う。m番目の近似モデルによる、μのx_iでの推定値(yのx_iでの予測値)は、次のように書ける。

$$\hat{\mu}_{mi} = \sum_{j=1}^{k_m} \hat{\theta}_j x_{ji},$$

ここで、$\hat{\theta}_j$は、X_{k_m}を、(i,j)の要素がx_{ij}である$n \times k_m$の行列で$Y = (y_1, \ldots, y_n)'$として、$\hat{\Theta}_m = (\hat{\theta}_1, \ldots, \hat{\theta}_{k_m})'$は

$$\hat{\Theta}_m = \left(X'_{k_m} X_{k_m}\right)^{-1} X'_{k_m} Y,$$

と定義できる。

モデル選択とモデル平均

先ほど見たとおり、いくつもの近似モデルを立てることができ、それらの違うモデルからもたらされる予測値は違ってくる。つまり、yの予測をしようとしても、その予測値としていくつもの違った値を考えることができる。

- モデル選択:　どれか一つのモデルを選んでその結果を使う。モデル選択はAICやBICといった情報量基準を立てて行われることが多い。多くの文献があるが、優れた教科書も多いのでそれらを参照されたい。

- モデル平均：　多くのモデルから出てきた結果を組み合わせて、予測を行う。この発表で取り扱う。教科書としては、Claeskens and Hjort (2008) など。

モデル平均

μ の M 個のモデルからの推定値を組み合わせる。$W = (w_1, \ldots, w_M)'$ をモデル平均のための重み付けベクトルとする。$W'\mathbf{1}_M = \sum_{m=1}^{M} w_m = 1$ という条件を課す。($\mathbf{1}_M$ は、$M \times 1$ の 1 からなるベクトルである。)

μ_i のモデル平均推定量は

$$\mu_i(W) = \sum_{m=1}^{M} w_m \hat{\mu}_{mi}$$

である。また、$P_m = X_{k_m}(X'_{k_m}X_{k_m})^{-1}X'_{k_m}$、$P(W) = \sum_{m=1}^{M} P_m$ とすると、次のようにも書ける。

$$\hat{\mu}(W) = P(W)Y = \sum_{m=1}^{M} w_m P_m Y.$$

重み付けベクトルの選び方

これまでの文献において、いろいろな重み付けベクトルの選び方が提唱されている。

- これまでの方法：　ベイズ統計を使った方法。情報量基準の大きさを使う方法。

- ここでは、Hansen (2007) にならい、（条件付）平均二乗誤差を最小化する。$\mu = (\mu_1, \ldots, \mu_n)'$かつ、$X = (x_1, \ldots, x_n)$とすると、

$$R_n(W) = E((\hat{\mu}(W) - \mu)'(\hat{\mu}(W) - \mu)|X).$$

平均二乗誤差の形

$R_n(W)$の式は、Hansen (2007)によって導出されている。$a_m = \eta_m'(I - P_m)\eta_m$
かつ、$\eta_m = (\eta_{m1}, \ldots, \eta_{mn})'$として、

$$A_n = \begin{pmatrix} a_1 & a_2 & a_3 & \ldots & a_M \\ a_2 & a_2 & a_3 & \ldots & a_M \\ a_3 & a_3 & a_3 & \ldots & a_M \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_M & a_M & a_M & \ldots & a_M \end{pmatrix}, \quad \Gamma_M = \begin{pmatrix} k_1 & k_1 & k_1 & \ldots & k_1 \\ k_1 & k_2 & k_2 & \ldots & k_2 \\ k_1 & k_2 & k_3 & \ldots & k_3 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ k_1 & k_2 & k_3 & \ldots & k_M \end{pmatrix}$$

とすると、

$$R_n(W) = W'(A_n + \sigma^2 \Gamma_M)W$$

となる。

最適な重み付けベクトル

最適な重み付けベクトルを$R_n(W)$を$W'\mathbf{1}_M = 1$の条件の下で最小化するものであるとする。

ここで、Wの要素を非負であるという制限をおかない。

- あるモデルの重みが負であるとは、そのモデルに含まれている変数を取り除くということと解釈できる。この性質によって、変数の順番付けに関する制約を少なくすることができる。

- Wの非負制約がないなら、最小化問題は、明示的な解をもち、計算が簡単になり、また最適な重み付けベクトルの解釈が容易になる。

Theorem 1.

$$W^* = (1'_M(A_n+\sigma^2\Gamma_M)^{-1}1_M)^{-1}(A_n+\sigma^2\Gamma_M)^{-1}1_M$$

$$= \begin{pmatrix} \frac{\sigma^2}{\sigma^2+\frac{a_1-a_2}{k_2-k_1}} \\ -\frac{\sigma^2}{\sigma^2+\frac{a_1-a_2}{k_2-k_1}}+\frac{\sigma^2}{\sigma^2+\frac{a_2-a_3}{k_3-k_2}} \\ -\frac{\sigma^2}{\sigma^2+\frac{a_2-a_3}{k_3-k_2}}+\frac{\sigma^2}{\sigma^2+\frac{a_3-a_4}{k_4-k_3}} \\ \cdots \\ -\frac{\sigma^2}{\sigma^2+\frac{a_{M-2}-a_{M-1}}{k_{M-1}-k_{M-2}}}+\frac{\sigma^2}{\sigma^2+\frac{a_{M-1}-a_M}{k_M-k_{M-1}}} \\ -\frac{\sigma^2}{\sigma^2+\frac{a_{M-1}-a_M}{k_M-k_{M-1}}}+1 \end{pmatrix}.$$

W^* の式からのいくつかの考察

- W^* の最初と最後の要素は常に正である。
- $W^* \in [0,1] \times [-1,1]^{M-2} \times [0,1]$.
- $(k_{m+1} - k_m)(a_{m-1} - a_m) > (k_m - k_{m-1})(a_m - a_{m+1})$ なら $W^*_m > 0$ である。もしこの条件が全ての m について満たされるなら、$W^* \in [0,1]^M$ となり、非負制約をおいた最適解と一致する。この条件は、変数の並べ方が適切である、と解釈できる。
- W^*_m は、$m+1$ 番目のモデルに入っている変数の説明力のみに依存し、それよりも高次のモデルに入っている変数には依存しない。

最適な重み付けベクトルの推定

先ほどみた最適な重み付けベクトルは、未知のパラメーターに依存しているので、データから計算できない。

ここでは、Hansen (2007) にならい、Mallows基準を最小化することによって、最適な重み付けベクトルを推定することにする。ここで、$\bar{e} = (\hat{e}_1, \ldots, \hat{e}_M)$, $e_j = Y'(I - P_j)Y$、かつ、$K = (k_1, \ldots, k_M)$ と定義する。最小化する基準は

$$W'\bar{e}'\bar{e}W + 2\sigma^2 K'W,$$

である。

この最小化問題は、ラグランジェ乗数法で簡単に解くことができる。

Theorem 2.

$$\hat{W} = -\frac{1}{2}(\bar{e}'\bar{e})^{-1}\left(2\sigma^2 K - (1'_M(\bar{e}'\bar{e})^{-1}1_M)^{-1}(2+2\sigma^2 1'_M(\bar{e}'\bar{e})^{-1}K)1_M\right)$$

$$= \begin{pmatrix} \sigma^2\frac{k_2-k_1}{\hat{e}'_1\hat{e}_1-\hat{e}'_2\hat{e}_2} \\ -\sigma^2\frac{k_2-k_1}{\hat{e}'_1\hat{e}_1-\hat{e}'_2\hat{e}_2}+\sigma^2\frac{k_3-k_2}{\hat{e}'_2\hat{e}_2-\hat{e}'_3\hat{e}_3} \\ -\sigma^2\frac{k_3-k_2}{\hat{e}'_2\hat{e}_2-\hat{e}'_3\hat{e}_3}+\sigma^2\frac{k_4-k_3}{\hat{e}'_3\hat{e}_3-\hat{e}'_4\hat{e}_4} \\ \cdots \\ -\sigma^2\frac{k_{M-1}-k_{M-1}}{\hat{e}'_{M-2}\hat{e}_{M-2}-\hat{e}'_{M-1}\hat{e}_{M-1}}+\sigma^2\frac{k_M-k_{M-1}}{\hat{e}'_{M-1}\hat{e}_{M-1}-\hat{e}'_M\hat{e}_M} \\ -\sigma^2\frac{k_M-k_{M-1}}{\hat{e}'_{M-1}\hat{e}_{M-1}-\hat{e}'_M\hat{e}_M}+1 \end{pmatrix}.$$

$\hat{W}$ の式からの考察

- $\hat{W}$ の最初の要素は常に正である。

- 前に $W^* \in [0,1] \times [-1,1]^{M-2} \times [0,1]$ であることを示したが、$\hat{W}$ はその集合に入っている保証はない。あとで見るように $\hat{W}$ を使うと実際にはうまくいかないのであるが、このことが、その問題の原因になっていると思われる。特に $\hat{e}'_m \hat{e}_m - \hat{e}'_{m+1} \hat{e}_{m+1}$ が小さいなら、対応する $\hat{W}$ の要素が非常に大きくなってしまう可能性がある。

- $(k_{m+1} - k_m)(\hat{e}'_{m-1} \hat{e}_{m-1} - \hat{e}'_m \hat{e}_m) > (k_m - k_{m-1})(\hat{e}'_m \hat{e}_m - \hat{e}'_{m+1} \hat{e}_{m+1})$ なら、m 番目の要素が正になる。もし、この条件が全ての m でみたされるの

なら、非負制約をおいた最適な重み付けベクトルの推定値と、おかないものとは一致する。

- また、$E(\hat{W}|X) \neq W^*$である。ただ、目的関数は不偏である。

- m番目の要素は、$m+1$番目までのモデルの残差にのみ依存し、それより高次のモデルの残差には依存しない。

yの予測値

yのxでの予測値を計算する。dを$n \times 1$のベクトルで、$x = X'_{k_M} d$となるものとする。するとモデル平均によるyの予測値は、

$$\hat{y} = d' X_{k_M} \hat{\Theta} = \sum_{m=1}^{M} w_m d' X_{k_m} \begin{pmatrix} \hat{\Theta}_m \\ 0 \end{pmatrix} = d' \sum_{m=1}^{M} w_m P_m Y = d' P(W) Y.$$

となる。

yの予測値の表現

$$
\begin{aligned}
P(\hat{W})Y &= \sigma^2 \frac{k_2 - k_1}{\hat{e}_1'\hat{e}_1 - \hat{e}_2'\hat{e}_2} P_1 Y + \left(1 - \sigma^2 \frac{k_M - k_{M-1}}{\hat{e}_{M-1}'\hat{e}_{M-1} - \hat{e}_M'\hat{e}_M}\right) P_M Y \\
&\quad + \sum_{m=2}^{M-1} \left(-\sigma^2 \frac{k_m - k_{m-1}}{\hat{e}_{m-1}'\hat{e}_{m-1} - \hat{e}_m'\hat{e}_m} + \sigma^2 \frac{k_{m+1} - k_m}{\hat{e}_m'\hat{e}_m - \hat{e}_{m+1}'\hat{e}_{m+1}}\right) P_m Y
\end{aligned}
$$

$\tilde{P}_m$を$\tilde{P}_1 = P_1$, $\tilde{P}_2 = P_2 - P_1, \ldots, \tilde{P}_M = P_M - P_{M-1}$として定義する。もし、回帰変数が互いに直行しているなら、$\tilde{P}_m$は、m番目のモデルではじめて登場した変数への射影行列である。すると

$$
\hat{e}_{m-1}'\hat{e}_{m-1} - \hat{e}_m'\hat{e}_m = Y'(I - P_{m-1})Y - Y'(I - P_m)Y = Y'\tilde{P}_m Y
$$

とかけるので、

$$
\hat{y} = d'P(\hat{W})Y = d'\tilde{P}_1 Y + \sum_{m=2}^{M} \left(1 - \sigma^2 \frac{k_m - k_{m-1}}{Y'\tilde{P}_m Y}\right) d'\tilde{P}_m Y.
$$

同等な方法

Mallows基準を最小化する重み付けベクトルを使ったモデル平均推定は、次の方法と同じである。

1. まず、m番目のモデルに入っていて$m-1$番目のモデルに入っていない変数を、$m-1$番目のモデルに入っている変数に回帰して、その残差をとることによって、変数の直交化を行う。

2. $\tilde{X}_m$を、先ほど直交化した、m番目のモデルに入っていて$m-1$番目のモデルに入っていない変数の行列とし、$\tilde{P}_m$をその射影行列とする。yをそれぞれの$\tilde{X}_m$に回帰する。すると、各変数の組に対応する係数推定値$\hat{\beta}_m = (\tilde{X}'_m\tilde{X}_m)^{-1}\tilde{X}'_m Y$を得る。

3. $\widehat{\beta}_m$ を

$$\widehat{\beta}^*_m = \left(1 - \sigma^2 \frac{k_m - k_{m-1}}{Y' \tilde{P}_m Y}\right) \widehat{\beta}_m$$

として、縮約する。James-Stein 推定量とは少し違うが非常によく似ている。
($k_m - k_{m-1} - 2$ を使わず、$k_m - k_{m-1}$ を使っている。)

4. $\tilde{X} = (\tilde{X}'_1, \ldots, \tilde{X}'_M)'$ での y の予測値を

$$\widehat{y} = \sum_{m=1}^{M} \tilde{X}'_m \widehat{\beta}^*_m$$

として計算する。

安定化

重み付けベクトルに非負制約を課さないモデル平均による推定量の性質を、安定的なものにするために、いくつかの方法を考えることができる。

- $W \in [0,1] \times [-1,1]^{M-2} \times [0,1]$ という制約をおいて、Mallows基準の最小化を行う。

安定化法その２

$\hat{W}$ を計算をするときに少しの perturbation を入れることによって、$\hat{W}$ の要素が非常に大きい値をとることがないようにする。つまり、ある小さな $\epsilon > 0$ を使って、

$$\hat{W}_\epsilon = \begin{pmatrix} \sigma^2 \frac{k_2 - k_1}{\hat{e}'_1 \hat{e}_1 - \hat{e}'_2 \hat{e}_2 + \epsilon} \\ -\sigma^2 \frac{k_2 - k_1}{\hat{e}'_1 \hat{e}_1 - \hat{e}'_2 \hat{e}_2 + \epsilon} + \sigma^2 \frac{k_3 - k_2}{\hat{e}'_2 \hat{e}_2 - \hat{e}'_3 \hat{e}_3 + \epsilon} \\ \cdots \\ -\sigma^2 \frac{k_{M-1} - k_{M-1}}{\hat{e}'_{M-2} \hat{e}_{M-2} - \hat{e}'_{M-1} \hat{e}_{M-1} + \epsilon} + \sigma^2 \frac{k_M - k_{M-1}}{\hat{e}'_{M-1} \hat{e}_{M-1} - \hat{e}'_M \hat{e}_M + \epsilon} \\ -\sigma^2 \frac{k_M - k_{M-1}}{\hat{e}'_{M-1} \hat{e}_{M-1} - \hat{e}'_M \hat{e}_M + \epsilon} + 1 \end{pmatrix}$$

とする。これは、Υ_M を (i, j) の要素が $M - \max(i, j)$ である $M \times M$ 行列として、

$$W'(\bar{e}'\bar{e} + \epsilon \Upsilon_M) W + \sigma^2 K' W,$$

を最小化することで得ることができる。

安定化法その３

ここでは、yの予測値を考える。それを

$$\widehat{y}^{+} = \sum_{m=1}^{M} \tilde{X}'_m \widehat{\beta}^{+}_m$$

として計算する。ここで、

$$\widehat{\beta}^{+}_m = \max\left(1 - \sigma^2 \frac{k_m - k_{m-1}}{Y'\tilde{P}_m Y}, 0\right) \widehat{\beta}_m$$

である。

この方法に対応する重み付けベクトルの形では、次のようになる。

$$
\begin{aligned}
\hat{w}_1^+ &= 1 - \max\left(1 - \sigma^2 \frac{k_2 - k_1}{Y'\tilde{P}_2 Y}, 0\right), \\
\hat{w}_m^+ &= \max\left(1 - \sigma^2 \frac{k_m - k_{m-1}}{Y'\tilde{P}_m Y}, 0\right) - \max\left(1 - \sigma^2 \frac{k_{m+1} - k_m}{Y'\tilde{P}_{m+1} Y}, 0\right), \\
\hat{w}_M^+ &= \max\left(1 - \sigma^2 \frac{k_M - k_{M-1}}{Y'\tilde{P}_M Y}, 0\right).
\end{aligned}
$$

なお、$\hat{W}^+ = (\hat{w}_1^+, \ldots, \hat{w}_M^+)' \in [0,1] \times [-1,1]^{M-2} \times [0,1]$である。

モンテカルロ実験

これまで見てきたモデル平均推定量の性質をモンテカルロ実験でみる。

まず、データは次のように生成する。

$$y_i = \sum_{j=1}^{M} \theta_j x_{ji} + e_i,$$

$i = 1, \ldots, N$であり、$x_{1i} = 1$で他の回帰変数は$x_{ji} \sim i.i.d.N(0, 1)$とする。誤差項は$e_i \sim i.i.d.N(0, 1)$から生成し、回帰変数とは独立である。

係数

係数については次の２つの特定化を考える。

- A: $\theta_j = cj^{-\alpha}$ for $j = 1, \ldots, M$

  cは定数である。αがどれぐらい高次の回帰変数が重要であるかを決める。

- B:

$$\begin{aligned} \theta_j &= 0, \text{ for } j = 1, \ldots, M/2, \\ \theta_j &= cj^{-\alpha} \text{ for } j = M/2 + 1, \ldots, M, \end{aligned}$$

  最初に、予測には役にたたない変数が入っていることに注意。つまり、変数の順序付けが間違っている。

推定量

- "P" 重み付けベクトルのすべての要素を正であるという制約をおく。
  "U" 重み付けベクトルに制約をおかない。
  "C" （安定化法１） $W \in [0,1] \times [-1,1]^{M-2} \times [0,1]$ という制約をおく。
  "R" （安定化法2） $W'(\bar{e}'\bar{e} + \epsilon\Upsilon)W + \sigma^2 K'W$ を最小化する。なお、$\epsilon = \hat{\sigma}^2$ とおく。
  "N" （安定化法3） β_m^+ あるいは w_m^+ を使う。

- 推定量の性質は、mean of the average squared error (MASE) で評価する:

$$E\left(\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{y}_i - \sum_{j=1}^{M}\theta_j x_{ji}\right)^2\right).$$

結果

- “U” はまったくうまくいかない。
- “R” は “U” よりうまくいく。“R” はϵの選び方が重要になってくる。
- “C” はうまくいっている。もし、回帰変数がすべて重要な場合 “P” よりも少し悪くなるものの、Mが比較的小さく、R^2が中間的な大きさのときで、回帰変数の並べ方が適切でないときには、“C” は “P” よりもうまくいっている。
- “N” はさらにうまくいっている。回帰変数が適切に並べられている場合でも、“P” と同じぐらいの誤差を返す。

結論

- 回帰変数が多い場合の線形回帰モデルのモデル平均推定を考えた。

- ここでは重み付けベクトルが非負であるという条件をおかない方法を考えた。これにより、変数の並べ方にそれほど依存しない方法を考えることができる。つまりモデル平均において、より多いモデルを少ない計算量で組み合わせることができるのである。

- モンテカルロ実験の結果から、“P” あるいは、“N” を実際には使用するべきであるといえる。どちらを使ったほうがよいかは、回帰変数の並べ方が適切であるかによって、変わってくる。