モデル平均理論の新展開

劉慶豊*
小樽商科大学

平成21 年7 月31 日

概 要

本論文はモデル選択とその発展系であるモデル平均(model averaging)に関する既存の研究結果を概観する。モデル平均の意味合いとモデル平均の応用及び未解決問題に関して説明する。さらに，モデル平均に関連する未解決問題に関する研究の将来的な方向を述べる。論文の中で実際のデータを利用した予測の例を示す。

1 はじめに

確率モデルをもとにデータ解析を行う際に，候補となるモデル群から一つ最良と思われるモデルを選ぶ手段として，モデル選択理論は盛んに研究されてきた。実証研究の主流として，AIC(Akaike [1973]）やBIC(Schwarz [1978])などの情報量基準を利用した選択法が広く利用されている。一方で，モデル選択に伴う不確実性を考慮して，ベイズ的な立場をとる研究者が異なったフィロソフィーに基づいて，ベイジアン・モデル平均の手法と理論を構築し発展させてきた。また，ここ数年，非ベイズ的な研究者は非ベイジアン・モデル平均理論を構築し始めた。本稿はモデル選択とモデル平均の理論を概観し，異なったフィロソフィーに基づく異なった手法の意味合いを明らかにする。さらに，この分野の未解決問題と研究における将来の発展方向の可能性について論じる。最後に実証例を挙げて，実証研究に適用するときの注意点を述べる。

第2章で，代表的な情報量基準であるAICやBICなどを紹介する。そして，第3章でモデル平均の意味合いを述べ，幾つかの例を挙げて，モデル平均の

*小樽商科大学商学部，Email:qliu@res.otaru-uc.ac.jp

推定量の性質に関して述べる。第4章で未解決問題に関して論じ，第5章でモデル平均を利用した幾つかの研究例を紹介した上で，実際のデータを用いて，ファイナンスへの応用を示す。第6章は結論である。

1.1 情報量基準によるモデル選択

統計学者や計量経済学者はモデルを構成して，データの特性をできるだけ正確に説明しようとする。一般的にはデータを生成する真の確率分布を近似するための候補として複数のモデルが考えられる。それらのモデルを評価するためにさまざまな情報量基準が提案されている。なかでも，AICやBICなどは最も広く利用されている。

1.2 赤池情報量基準AIC

AICの基本となっているものはカルバック－ライブラー距離である。確率変数yの標本$\mathbf{y} = (y_1, y_2, \cdots, y_n)$が観測されたとする。$y$の真の密度関数を$g(y)$としてパラメットリックなモデル$f(y, \boldsymbol{\theta})$で近似しようとする。ただし，$\boldsymbol{\theta} \in \Theta$で$\Theta \subseteq \mathbb{R}^K$である。$f(y, \boldsymbol{\theta})$による近似の良さを計るための一つの基準として，Kullback and Leibler [1951]がカルバック－ライブラー距離を提案した。その定義は以下のようになる，

$$\begin{aligned} KL(g, f) &= E_g\left(\log \frac{g(y)}{f(y, \boldsymbol{\theta})}\right) \\ &= \int_{-\infty}^{\infty} \{\log g(y) - \log f(y, \boldsymbol{\theta})\} g(y)\, dy. \end{aligned} \tag{1}$$

$\log(\cdot)$は自然対数を表す。カルバック－ライブラー距離はモデルで真の分布を近似するときの一種のリスク関数として考えられる。

$KL(g, f)$の中の$g(y)$はモデルに依存しないため，モデルの良さを評価する場合，$\log g(y)$の部分を無視できる。原理的には平均対数尤度と呼ばれる

$$E_g(\log f(y, \theta)) = \int_{-\infty}^{\infty} \log f(y, \boldsymbol{\theta})\, g(y)\, dy \tag{2}$$

を最大にするようなモデルが最適なモデルとなる。ここで，大数の法則を利用すれば

$$\lim \frac{1}{n} \sum_{i=1}^{n} \log f(y_i, \boldsymbol{\theta}) = E_g(\log f(y, \boldsymbol{\theta})) \tag{3}$$

が分かる。そのため，対数尤度関数を

$$l_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i, \boldsymbol{\theta}) \tag{4}$$

と定義し，$\hat{\boldsymbol{\theta}}$を最尤推定量として，$\frac{1}{n} l_n(\hat{\boldsymbol{\theta}})$が平均対数尤度$E_g(\log f(y, \hat{\boldsymbol{\theta}}))$の一つの推定量として考えられるが，この推定量はバイアスを持っていることが知られている（証明は小西・北川 [2004]を参照されたい)。そこで，バイアスを修正して，モデル選択の基準としてAICが構築された。AICは

$$\mathrm{AIC} = -2 l_n(\hat{\boldsymbol{\theta}}) + 2p \tag{5}$$

と定義される。ただし，pはパラメーター$\boldsymbol{\theta}$の次元で，平均対数尤度を推定するときのバイアスの補正として働く。$2p$はパラメーターの数，すなわちモデルの複雑さに対するペナルティとして解釈されることもある。モデルを選択する際，AICを最小にするモデルを選ぶ。

1.3 他の情報量基準

AICと並んでベイジアン情報量基準(BIC）も広く利用されている。他に，AICの修正版として竹内 [1976]とStone [1977]によって考案された情報量基準TICもよく知られている。

BICの背景となっているのはベイジアン統計学である。事後確率$P(M|\mathbf{y})$が最大になるようなモデルMを選ぶというのがその発想である。導出の詳細はClaeskens and Hjort [2008]を参照されたい。事後確率の近似的な評価値としてBICが定義される。

$$\mathrm{BIC} = -2 l_n(\hat{\boldsymbol{\theta}}) + (\log n)\, p \tag{6}$$

AICと似た形になっているが，そのペナルティ$(\log n)\, p$はnが大きくなるときAICのペナルティより大きくなる。nが大きくなるに連れAICはより複雑なモデルを選びがちであるが，BICは真のモデルよりパラメーター数の少ないモデルを選ぶ傾向がある。そして，真のモデルが候補となるモデル族の中に入っていれば，サンプル数nが無限大に近づくにつれ，BICが真のモデルを選ぶ確率は1に収束する。すなわち，BICはモデル選択の情報量基準として一致性を持つ。その一方では，AICは一致性を持たない。

小西・北川 [2004]によれば、AICを計算する際、モデルがデータを生成する確率分布と一致するときに限って，pが平均対数尤度を推定するときのバイ

アスの補正項になることが保障される。そうではないとき，pが正しいバイアスの補正項とならない。竹内 [1976]は正確な補正項を導出して，想定したモデル$f\left(y,\boldsymbol{\theta}\right)$の中に真の確率分布を含まれていない場合で利用できる情報量基準(TIC）を構築した。

$$\mathrm{TIC}=-2l_n(\hat{\boldsymbol{\theta}})+2Tr\left(\hat{I}\hat{J}^{-1}\right) \tag{7}$$

$\hat{I}$と$\hat{J}$はそれぞれ

$$I_0=E_g\left[\frac{\partial\log f\left(y,\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\frac{\partial\log f\left(y,\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}'}\right] \tag{8}$$

$$J_0=-E_g\left[\frac{\partial^2\log f\left(y,\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] \tag{9}$$

の推定量で，一致推定量$\hat{\boldsymbol{\theta}}$を$I_0$と$J_0$のサンプル版に代入することで得られる。モデルはデータを生成する確率分布と一致しないとき，AICが厳密的には正確な補正項を使っていないが，想定するモデルが真の確率分布に近い場合，近似的にはAICをモデル選択の手法として利用することが可能で，実際にも広く使われている。

また，線形回帰モデルの情報量基準としてMallows [1973]はMallows'C_p

$$C_p=RSS+2p\hat{\sigma}^{2}-n\hat{\sigma}^{2} \tag{10}$$

を定義した。C_pはリスク関数Mean Squared Error(MSE)の推定量となっている。それを最小にするように，モデルに入れる説明変数を選ぶ。C_pは線形モデルにしか適用できないが，非線形モデルは線形化してからの適用が可能である。

2　モデル平均に関して

実証研究で，AICやBICなどの情報量基準を利用してモデル選択を行う際，選ばれたモデルをそのまま真のモデルとして推論が展開されることが多い。しかし，AICやBICなどの情報量基準によるモデル選択は一定の不確実性をもたらす。その不確実性が推定量の性質に影響を及ぼすことを考慮しないと間違った推論に繋がる。そして，その推論のもとで計算した信頼区間や行った検定の結果が誤ったものになる。この様な問題が選択後推定量（Post Model

Selection Estimator，PMSE）の問題と呼ばれる。モデル平均はこのようなPMSEの問題を考慮に入れた手法である。

モデル平均はモデル選択を拡張して得られた手法である。候補となるモデル族を$\mathcal{M} = \{M_1, M_2, \cdots, M_K\}$とする。ある母数$\mu$のモデル平均による推定量の一般表現は

$$\hat{\mu} = \sum_{M_k \in \mathcal{M}} c\left(M_k\right) \hat{\mu}_{M_k} \tag{11}$$

で与えられる。$\hat{\mu}_{M_k}$はモデルM_kのもとで得られる推定量で，$c\left(M_k\right)$は各モデルに与えられるウエイトを決める関数である。$c\left(M_k\right)$は総和が1で，非確率的でも確率的でも良い。この表現は情報量基準でモデルを選択して得られる推定量を含んでいる。たとえば，AICの場合，AICで選んだモデルをM_{AIC}とする。そのモデルによって得られる推定量は

$$\hat{\mu}_{AIC} = \sum_{M_k \in \mathcal{M}} I\left(M_k = M_{AIC}\right) \hat{\mu}_{M_k} \tag{12}$$

となる。一つ注意すべきは，AICを基準にモデルを選ぶ際，どのモデルが選ばれるかは確率的なイベントであり，$I\left(M_k = M_{AIC}\right)$が確率変数となっていることである。後ほどこの点に関して詳しく論じることにする。

2.1　モデル選択はモデル平均の特殊ケース

AICで選んだモデルよって得られた推定量の場合，最小のAICに対応しているモデルにウエイト1を与えて，残りのモデルに0のウエイトを与えている。モデル平均による推定量の場合，ウエイトはより一般的で，AICの場合のウエイトを特殊ケースとして含む。そのため，リスクで評価するとき，リスクに関する最適なウエイトを使えば，必ずモデル平均による推定量のリスクは，AICの場合のものと等しいかより小さくなる。この意味ではモデル平均の手法を利用することで、単純にAICなどの情報量基準による手法より優れた結果を得る可能性がある。他の特定の一つのモデルしか選ばない伝統的なモデル選択の手法に関しても同様のことが言える。

$\hat{\mu}_{AIC}$の例や，モデル平均による推定量の一般表現からも分かるように，推定量の性質を導出する際，ウエイトとなる部分の確率的な特性による影響，すなわち，モデル選択に伴う不確実性を考慮しなければならない。例えば，$\hat{\mu}_{AIC}$は本質的にM_{AIC}が真のモデルであるという条件のもとで得られた推定量と異なるものである。そのため，AICでモデルM_{AIC}を選択し，選択されたモデル

を真のモデルとして推定量の性質を導出するのは誤ったやり方である。これはいわゆるモデル選択後推定量（Post-Model-Selection Estimator，PMSE）の問題である。これから，いくつかのモデル平均の手法を紹介してから，PMSEの問題を論じることにする。

2.2 ベイジアンモデル平均

近年，ベイズ的な枠組みでモデル平均の手法に関する論文が数多く発表されてきた（Draper [1995]，Hoeting et al. [1999]，Clyde and George [2004]）。ベイジアンの枠組みでは，モデルのパラメーターを確率変数として考える。関心のある母数をμとして，各モデルのパラメーター$\boldsymbol{\theta}_k$の関数とする。その母数の事後分布$\pi(\mu|\mathbf{y})$や条件付期待値$\hat{\mu} = E(\mu|\mathbf{y})$を推定することが目的である。

これまでの設定と同じく，候補となるモデル族を$\mathcal{M} = \{M_1, M_2, \cdots, M_K\}$とする。ベイズの公式を利用して，モデル平均によって$\pi(\mu|\mathbf{y})$と$E(\mu|\mathbf{y})$が導出される。そのためには，二つの事前分布が必要となる。一つは各モデルの事前分布で$P(M_k)$と記す。もう一つは各モデルのパラメーター$\boldsymbol{\theta}_k$のそのモデルのもとでのパラメーターの事前分布$\pi(\boldsymbol{\theta}_k|M_k)$である。この二つの事前分布を所与として，$\mathbf{y}$のモデル$M_k$のもとでの尤度関数を$L(\mathbf{y}|M_k, \boldsymbol{\theta}_k)$として，モデル$M_k$の事後分布は

$$P(M_k|\mathbf{y}) = \frac{P(M_k)\lambda_k}{\sum_{k=1}^{K} P(M_k)\lambda_k} \tag{13}$$

となる。ただし

$$\lambda_k = \int L(\mathbf{y}|M_k, \boldsymbol{\theta}_k)\, \pi(\boldsymbol{\theta}_k|M_k)\, d\boldsymbol{\theta}_k \tag{14}$$

である。さらにμの事後分布は

$$\pi(\mu|\mathbf{y}) = \sum_{k=1}^{K} P(M_k|\mathbf{y})\, \pi(\mu|M_k, \mathbf{y}) \tag{15}$$

となる，ただし，$\pi(\mu|M_k, \mathbf{y})$はモデルM_kが真のモデルである条件のもとでのμの事後分布であり，事前分布$\pi(\boldsymbol{\theta}_k|M_k)$を利用してベイズの公式により解析的に，またはMCMCなどの方法で数値的に推定できる。そしてμの事後期待値は

$$\hat{\mu} = E(\mu|\mathbf{y}) = \sum_{k=1}^{K} P(M_k|\mathbf{y})\, E(\mu|M_k, \mathbf{y}) \tag{16}$$

と導出される。

ベイジアンモデル平均の手法はすでに多くの成果をあげているが，問題点も残されている。たとえば候補となるモデルに与える事前確率を決めるとき，かなり主観的で場当たりな方法が取られる。それにより推定結果が大きく左右される。

2.3 Smoothed-AIC，BIC

Buckland et al. [1997]はモデルのAICおよびBICの値を利用したモデル平均を方法を提案した。この方法はベイジアンモデル平均と密接な関連を持っている。

前述したようにBICは事後確率$P(M_k|\mathbf{y})$を評価している。Claeskens and Hjort [2008]によれば近似的に

$$\mathrm{BIC} \approx -2\log(\lambda_k) \tag{17}$$

が成立する。そこで$P(M_k)$がkに関して均一であると仮定すれば，(13) 式から

$$P(M_k|\mathbf{y}) \approx \frac{\exp(-\mathrm{BIC}_k/2)}{\sum_{k=1}^{K}\exp(-\mathrm{BIC}_k/2)} \tag{18}$$

であることがわかる。(16)式はモデル平均のウエイトとして

$$c_{BIC}(M_k) = \frac{\exp(-\mathrm{BIC}_k/2)}{\sum_{k=1}^{K}\exp(-\mathrm{BIC}_k/2)} \tag{19}$$

が一つ妥当な選択肢であることを示唆する。そのウエイトを利用して構築したモデル平均の推定量を

$$\hat{\mu}_{MA-BIC} = \sum_{M_k\in\mathcal{M}} c_{BIC}(M_k)\,\hat{\mu}_{M_k} \tag{20}$$

と定義できる。この推定量はsmoothed-BIC-based estimatorと呼ばれる。

$\hat{\mu}_{MA-BIC}$の定義と同じフォームを取って，smoothed-AIC-based estimatorが

$$\hat{\mu}_{MA-AIC} = \sum_{M_k\in\mathcal{M}} c_{AIC}(M_k)\,\hat{\mu}_{M_k} \tag{21}$$

と定義される。ただし，

$$c_{AIC}(M_k) = \frac{\exp(-\mathrm{AIC}_k/2)}{\sum_{k=1}^{K}\exp(-\mathrm{AIC}_k/2)} \tag{22}$$

である。モデル族$\mathcal{M}$に含まれる各モデルに最尤法を適用すれば，各モデルの推定量とウエイトを計算できるため，この二つのモデル平均の推定量の計算は非常に簡単である。

2.4 Mallows'C_pを用いたモデル平均

Hansen [2007]はMallows'C_pを利用したモデル平均の方法を提案した。モデルは線形モデルとして表現されている。モデル式は

$$y_i = \mu_i + \varepsilon_i \tag{23}$$

$$\mu_i = \sum_{j=1}^{\infty} \theta_j x_{ji} \tag{24}$$

$$E\left(\varepsilon_i | x_i\right) = 0 \tag{25}$$

$$E\left(\varepsilon_i^2 | x_i\right) = \sigma^2 \tag{26}$$

である。$(y_i, \mathbf{x}_i)$，$\mathbf{x}_i = (x_{1i}, x_{2i}, \cdots)$，$i = 1, 2, \cdots, n$はランダムなサンプルで，$\varepsilon_i$が撹乱項である。$\sum_{j=1}^{\infty} \theta_j x_{ji}$を非線形またはノンパラメトリックな関数の級数展開として考えることができるので，この方法の適用範囲は線形モデルだけに限定されない。

このモデルのもとで，$\mathbf{x}_i$の最初のk個の要素だけを説明変数とするモデル

$$\mu_i = \sum_{j=1}^{k} \theta_j x_{ji} \tag{27}$$

をM_kで表す。最小のモデルがM_1で，最大のモデルをM_Kと限定する。モデル平均の推定量は

$$\hat{\boldsymbol{\theta}} = \sum_{k=1}^{K} \omega_k \begin{pmatrix} \hat{\boldsymbol{\theta}}_k \\ \mathbf{0} \end{pmatrix} \tag{28}$$

と定義される。ただし，$\hat{\boldsymbol{\theta}}_k = (\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k)'$は$\boldsymbol{\theta}_k$のモデル$M_k$のもとでの推定量で，$\boldsymbol{\theta}_k = (\theta_1, \theta_2, \cdots, \theta_k)'$である。$\omega_k$はモデル$M_k$に与えるウエイトで，$\omega_k \in [0, 1]$, $\sum_{k=1}^{K} \omega_k = 1$を満たす。$K$個のウエイトをまとめて$W = (\omega_1, \omega_2, \cdots, \omega_K)'$と記す。

$\mathbf{X}$を$\mathbf{x}_i, i = 1, 2, \cdots, n$,のデータ行列とする。$\hat{\mu}(W) = \mathbf{X}\hat{\boldsymbol{\theta}}$は$W$をウエイトとして利用して得られた$\mu$のモデル平均の推定量を表す。損失関数を$L_n(W) =$

$(\hat{\mu}(W)-\mu)'(\hat{\mu}(W)-\mu)$として，Hansen [2007]はMallows'C_pが一定の条件のもとで，損失関数の期待値と定数の和の不偏推定量であることを示した。ただし，Mallows'C_pの定義は行列表現で

$$C_n(W)=\left(Y-X_K\hat{\Theta}\right)'\left(Y-X_K\hat{\Theta}\right)+2\sigma^2k^* \tag{29}$$

となる。定義式の中で，X_Kは$k=K$の場合のモデル，すなわち最大のモデルの説明変数のデータ行列で，$k^*=\sum_{k=1}^{K}\omega_k k$である。さらに，Mallows'$C_p$，$C_n(W)$を最小にするウエイトを

$$\hat{W}_n=\arg\min_W C_n(W) \tag{30}$$

と定義する。一定の条件のもとで

$$\frac{L_n\left(\hat{W}_n\right)}{\inf_W L_n(W)}\xrightarrow{p}1 \tag{31}$$

が満たされ，$\hat{W}_n$は漸近的に損失関数の値を最小にする最適なウエイトとなる。

2.5 PMSE推定量の漸近分布

モデル選択で選んだモデルのもとで得られる推定量，およびモデル平均による推定量は，Post-model-selection(PMSE)推定量と呼ばれている。このように呼ぶのは，それらの推定量はモデル選択やモデル平均のステップによってその漸近的な性質が影響され，推定したモデルが真のモデルであるという仮定のもとでの推定量とは異なることを表現したいためである。モデル選択とモデル平均によってもたらされる不確実性を考慮に入れ，PMSE推定量の漸近分布の導出を行った論文は幾つか存在するが（Hjort and Claeskens [2003], Potscher [2006]とLeeb and Potscher [2008]），ここでは，Hjort and Claeskens [2003]のアプローチに関して簡単に紹介する。

Hjort and Claeskens [2003]はlocal misspecificationのもとで漸近分布の導出を行った。（Potscher [2006]とLeeb and Potscher [2008]も同様に，local misspecificationのもとで漸近分布を導出している。）真のモデルをサンプル数に依存する密度関数の列で定義している。$y_i, i=1,2,\cdots,n$が確率変数yのサンプルで，その密度関数は

$$f_{true}(y)=f_n(y)=f\left(y,\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0+\boldsymbol{\delta}/\sqrt{n}\right) \tag{32}$$

とする。$\boldsymbol{\theta}_0$と$\boldsymbol{\gamma}_0$は次元がpとKのパラメーターベクトルである。最小のモデルは$\boldsymbol{\gamma}$に属するパラメーターを含まない，$f_{narr}\left(y,\boldsymbol{\theta}\right)=f\left(y,\boldsymbol{\theta},\boldsymbol{\gamma}_0\right)$である。最大のモデルはすべてのパラメーターを含むようなモデルで$f_{full}\left(y,\boldsymbol{\theta},\boldsymbol{\gamma}\right)$と表す。$\boldsymbol{\gamma}$に属するパラメーターの中から幾つかを選んでモデルに入れることによって，最大と最小のモデルの中間にあるサブモデルを構築することができる。M_{S_j}でサブモデルを表す。S_jはモデルM_{S_j}に入れた$\boldsymbol{\gamma}$の要素の添え字の集合である。$\boldsymbol{\gamma}$の要素のすべての組み合わせを考えると，最大と最小のモデルを含めて，サブモデルは全部で2^K種類がある。ある母数μを推定したい場合，そのモデル平均推定量は

$$\hat{\mu}=\sum_{j\in 2^K} c\left(M_{S_j}\right)\hat{\mu}_{S_j} \tag{33}$$

と定義される。ただし，$\hat{\mu}_{S_j}$はモデルM_{S_j}のもとでの推定量である。

Hjort and Claeskens [2003]はこのような設定のもとで，最尤法のテクニックを駆使してモデル平均推定量$\hat{\mu}$の漸近分布を導出した。一般的には，$c(M_{S_j})$が確率変数となり，導出された分布は標準的な分布に従わず，一種の正規分布の混合分布となる。そして，モデル選択による推定量の漸近分布はその特殊ケースとして含まれる。Hjort and Claeskens [2003]は漸近分布の結果を利用して，モデル選択で得られたモデルを真のモデルとして，信頼空間を導出した場合，信頼空間が過小評価されることを論じた。また，幾つかの異なったモデル平均推定量のリスクの比較を行った。

3　未解決問題

統計的推論に関しては前節で紹介したPMSE推定量の分布を利用して推定量の有限標本の分布を近似し，検定や信頼区間の導出を行うのは一つの自然な考え方である。しかし，このような近似はサンプル数の大きさに関わらず，必要な精度に到達するのは不可能であることがPotscher [2006]やLeeb and Potscher [2008])などにより示されている。

以下ではLeeb and Potscher [2008]の結果を説明する。モデルを行列表現で

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\lambda}+\boldsymbol{\varepsilon} \tag{34}$$

と表す。$\mathbf{X}$と$\mathbf{Z}$はそれぞれ$(n\times k)$と$(n\times q)$の非確率的な説明変数の行列である。$\boldsymbol{\varepsilon}\sim N\left(0,\sigma^2\mathbf{I}_n\right)$は撹乱項である。パラメーターをまとめて$\boldsymbol{\theta}=\left(\boldsymbol{\beta}',\boldsymbol{\lambda}'\right)'$とする。$\mathbf{X}$に含まれる説明変数は必ずモデルに入れることにする。モデル選択によ

り，$\mathbf{Z}$に含まれる説明変数のうち，どれを残すかを決める。そして，選択されたモデルのもとで，$\boldsymbol{\beta}$を含むパラメーターを推定する。$\boldsymbol{\beta}$の推定量$\hat{\boldsymbol{\beta}}$の有限標本の分布を$G_{n,\boldsymbol{\theta},\sigma}(\mathbf{b})$とする。Leeb and Potscher [2008]によれば，$G_{n,\boldsymbol{\theta},\sigma}(\mathbf{b})$の推定量として$\hat{G}_n(\mathbf{b})$が任意の正の数$\delta$に関して

$$P_{n,\boldsymbol{\theta},\sigma}\left(\left|\hat{G}_n(\mathbf{b}) - G_{n,\boldsymbol{\theta},\sigma}(\mathbf{b})\right| > \delta\right) \overset{n\to\infty}{\to} 0 \tag{35}$$

を満たすような，すなわち$\boldsymbol{\theta}$に関してポイントワイズに一致性を持つ推定量を構築することができるが，一方で，ある正の数ηが存在して，その推定量$\hat{G}_n(\mathbf{b})$について

$$\sup_{\|\boldsymbol{\theta}\|<\eta} P_{n,\boldsymbol{\theta},\sigma}\left(\left|\hat{G}_n(\mathbf{b}) - G_{n,\boldsymbol{\theta},\sigma}(\mathbf{b})\right| > \delta\right) \overset{n\to\infty}{\to} 1 \tag{36}$$

が成り立つ。これは$\hat{G}_n(\mathbf{b})$はパラメーター$\boldsymbol{\theta}$に関して一様な一致性を持たないことを意味する。Leeb and Potscher [2008]は，この一様に一致性を持つ推定量が存在しないという結論に不可能性と名付けた。不可能性は以上のモデルの設定のもとだけで成立するのではなく，より一般的なパラメトリックとセミパラメトリックモデルに関しても成立する。さらに，不可能性は殆どのモデル選択の方法，そしてモデル平均に関しても成り立つ。

不可能性の結論により，サンプル数を如何に大きくしても，PMSE推定量の有限標本の分布の推定がうまく行かないことが分かる。そのため，有限標本分布の推定量のもとで推論を行うアプローチが実現できない。そこで，Beran and Dumbgen [1998]とBeran [2000]などはこの問題を解決するための一つの方向を提示している。その大まかな方針は以下のようになっている。まず，候補となるモデルに対応するリスクを推定する。そして，リスクを最小にするモデルを選ぶ。次は，選んだモデルのリスクの推定量の漸近分布を導出する。最終的にはリスクの推定量の漸近分布のもとで統計的推論を行う。Beran and Dumbgen [1998]とBeran [2000]はこのアプローチに基づいて推定量を中心とした母数の信頼集合を導出した。

4　モデル平均の応用

モデル選択の手法は各分野で広く利用されている。それとは対照的に，モデル平均の手法の応用はまだ数が少ない。これから，幾つかのモデル平均を利用した研究を紹介する。その上，実際のデータを利用して，モデル平均による予測の例を示す。

4.1 既存研究

経済理論を計量経済学の手法を利用して分析するとき，モーメント条件を利用して推定や推論を行うのは一つ重要な手法である。このような手法は一定の数のモーメント条件しかない場合，基本的に容易に利用できるが，モーメントの数が数多く存在する場合，すべてのモーメント条件を利用することは殆どの場合推定量の性質を低下させるため，モーメント条件の選び方を考える必要がある。この問題に関しては，Morimune [1983]は一つ重要な研究となっている。そして，森棟公夫 [1985]，Bekker [1994]，Donald and Newey [2001]，Kuersteiner and Okui [2009]などはさらにその問題に関して研究を進めてきた。そして，Kuersteiner and Okui [2009]はモーメントが複数存在するという問題の特殊ケースである操作変数の選択の問題にモデル平均の手法を応用した。Kuersteiner and Okui [2009]は二段階最小二乗法の第一段階でHansen [2007]が提案したモデル平均の手法を利用し操作変数の選択を行い，最終的には最小の損失関数の値をもたらす推定量を導出した。

モデル平均の手法をボラティリティモデルの推定やValue-at-Rsikの分析に応用した実証研究が幾つか存在する。Brownlees and Gallo [2008]はSmoothed-AIC，BIC，FIC[1]の手法を用いてボラティリティモデルの変数選択を行った。また，Pesaran et al. [2009]はモデル平均の手法を利用してValue-at-Rsikの分析を行い，モデル平均の推定結果に利用できる診断テストを考案した。

4.2 モデル平均による予測

これから日本の株価指数のデータを利用して，モデル平均の一つの適用例を示す。日本の日経225のデータに関して，幾つかの候補となるARCH(Autoregressive Conditional Heteroscedasticity)型モデルを推定して，株価指数の実現ボラティリティの予測を行う。

ARCH型モデルは株価のボラティリティを分析するための最もポピュラーなモデルのクラスのひとつである。Engle [1982]によって時系列データの分散不均一性を捉えるためのモデルとしてARCHモデルが考案されてから，データの特性を様々な側面から捉えるため，数多くのバラエティが提案されてきた。ここでは，その中の幾つかの代表的なモデルをモデル平均の候補モデルとする。以下では，簡単にそれらの候補モデルを紹介する。

[1]FICはFocused Information Criterionの略である。推定対象となる異なった母数の特質に配慮した情報量基準である。詳細に関してはHjort and Claeskens [2003]またはClaeskens and Hjort [2008]を参照されたい。

4.2.1 ARCH型モデル

Engle [1982]は時系列データの分散不均一性を捉えるためのモデルとして，ARCH(q)モデルを提唱した。ARCH(q)モデルにおいては，条件付分散が不均一であることが仮定されている。$t-1$期までの情報集合ψ_{t-1}に含まれる内生変数のラグと外生変数の線形結合$\mathbf{x}_t\beta$によって確率変数r_tの条件付平均があらわされる。また，その条件付分散をh_tとする。そのモデル式は

$$r_t = \mathbf{x}_t\boldsymbol{\beta} + \sqrt{h_t}z_t \quad or \quad r_t|\psi_{t-1} \sim N[\mathbf{x}_t\boldsymbol{\beta}, h_t] \tag{37}$$

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 \tag{38}$$

$$\varepsilon_t = \sqrt{h_t}z_t \quad z_t \sim N[0,1]\ i.i.d.. \tag{39}$$

となる。モデル式の中のh_t，場合によっては$\sqrt{h_t}$は，ファイナンスモデルにおけるボラティリティに該当する。モデル式を見れば分かるように，ARCH(q)モデルはボラティリティが過去の予期せぬリターンのショックの二乗によって影響されることを表現している。h_tは条件付分散なので，非負でなければならない。従って，$\alpha_0 > 0, \alpha_i \geq 0$という制約が必要となる。

実務上，ARCH(q)モデルを金融データに適用して，ボラティリティに対するショックの持続性を表現しようとすると，ARCH(q)モデルの次数が長くなるという問題が生じる。この問題を回避するために提案されたのはGARCH(p,q)モデル(Bollerslev [1986])である。GARCH(p,q)モデルはARCH(q)モデルにボラティリティのラグの項を追加したものである。GARCH(p,q)モデルのボラティリティに関する式は次のようになる。

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j} \tag{40}$$

ARモデルにMA項を追加することで，ARモデルの次数を減らせることと同様に，GARCHモデルにおいてはボラティリティのラグ項を導入することで，短い次数，普通はGARCH(1,1)モデルで，データの特性を捉えることができるようになる。

GARCHモデルにおいても，ボラティリティの非負性を保つため，各パラメーターに制約をおく必要がある。ARCHモデルにおける制約

$$\alpha_0 > 0, \alpha_i \geq 0 \quad i = 1,2,\cdots,q$$

に加えて，$\beta_j \geq 0, j = 1,2,\cdots p$の制約条件が必要となる。

実証研究を積み重ねるうちに，金融データのボラティリティに関する様々な特徴が明らかになってきた。その中の一つとして，ボラティリティ変動の非対称性があげられる。非対称性とは，前日の株価変動の方向の違いによって，ボラティリティの増減の大きさの具合が変わるという特性である。経験的には前日の株価が下落した場合には，前日の株価が上昇した場合よりも，今日のボラティリティは大きくなる傾向が強い。しかし，GARCHモデルでは，この非対称性を捉えることができない。そこで，GJR-GARCHモデル，EGARCHモデル，TGARCHモデルなどが提案された。

Nelson [1991]によって提案されたEGARCH(Exponential GARCH)モデルは非説明変数のh_tを$\log(h_t)$に置き換えることによって，ボラティリティ変動の非対称性を捉えることを可能にし，またGARCHモデルにおけるパラメーターの非負制約を除去することにも成功した。EGARCH(p,q)モデルは以下の式で表される。

$$\log(h_t) = \alpha_0 + \sum_{i=1}^{q} \alpha_i[\theta z_{t-i} + \gamma(|z_{t-i}| - E|z_{t-i}|)] + \sum_{j=1}^{p} \beta_j \log(h_{t-j}) \qquad (41)$$

$$\varepsilon_t = \sqrt{h_t} z_t \quad z_t \sim N[0,1] \ i.i.d. \qquad (42)$$

ここで，$\theta < 0$であれば，前述した経験的事実のボラティリティ変動の非対称性が表現されることになる。$\ln(h_t)$がマイナスの値をとることもできるので，パラメーターに関する非負制約が不要となる。ただし，EGARCHモデルには欠点もある。GARCHモデルと異なり，ε_tの代わりにz_tでモデルを定式化したため，特定のショックε_tのボラティリティの変動に与える影響がわかりにくくなるのである。

Liu and Morimune [2006]は株価の連続的な上昇または下落がボラティリティに与える影響を捉えるため，OGARCHモデルを提案した。OGARCH(1, 1)モデルのボラティリティ式は

$$h_t = \alpha_0 + \alpha_1 \exp(\phi\gamma_{t-1})\varepsilon_{t-1}^2 + \beta h_{t-1}, \qquad (43)$$

となる。α_0， α_1，βとϕは未知パラメーターで，$a_0 \geq 0$，$a_1 > 0$，$\beta > 0$である。そして，γ_{t-1}は第t期まで同じ符号のショックが連続的に現れた日数である。その定義は

$$\gamma_{t-1} \equiv i, \text{ if } sign(\varepsilon_{t-1}) = \cdots = sign(\varepsilon_{t-i}) = -sign(\varepsilon_{t-(i+1)}) \qquad (44)$$

である。γ_{t-1}は1か$\gamma_{t-2} + 1$という値をとる。モデル式の中の指数関数の項は非負で，係数ϕが正の値と期待される。すなわち，株価が同じ方向に連続的に動くと，ボラティリティが大きくなりがちであると期待する。

4.2.2　実現ボラティリティの予測

応用例では，GARCH(1,1)，OGARCH(1,1)及びEGARCH(1,1)モデルを利用する。モデル平均はSmoothed BICを採用する。データは1984年1月9日から2007年12月25日までの日経225の日次データと週次データである。[2]日次と週次データのサンプルサイズはそれぞれ5902と1242である。日次データから週次の実現ボラティリティ（Realized Volatility，RV)を計算する。そして，週次データで上記ARCH型モデルを推定し，RVを予測する。週次の株価をy_t，$t = 1, 2, \cdots n$として，ログ・リターンは

$$r_t = 100 \left[\log \left(y_t \right) - \log \left(y_{t-1} \right) \right] \tag{45}$$

とする。週次のRVを

$$RV_t = \sum_{i=1}^{n_t} r_{t_i}^2 \tag{46}$$

と定義する。ただし，n_tは第t週間にあった営業日の数で，r_{t_i}はその週の第i番目のログ・リターンとする。ローリング・ウインドの幅を442とする。442週間のデータを利用してモデルを推定し，一週間先のRVを予測する。ローリング・ウインドを移動しながら，残りの800週間のRVを予測する。予測の良さを計るため，平均予測誤差二乗和

$$MSPE = \frac{1}{n} \sum_{t=1}^{n} \left(\widehat{RV}_t - RV_t \right)^2 \tag{47}$$

を利用する。ただし，$\widehat{RV}_t$はRV_tの推定値を表す。二通りのSmoothed BICによるモデル平均を行い，予測する。一つはGARCH(1,1)とOGARCH(1,1)との二つのモデルの平均でもう一つはGARCH(1,1)，OGARCH(1,1)とEGARCH(1,1)との三つのモデルによる平均である。モデル平均による予測値は

$$\widetilde{RV}_t = \sum_{m=1}^{M} c_{BIC} \left(m \right) \widehat{RV}_{m,t} \tag{48}$$

と定義する。ただし，Mはモデル平均に使われるモデルの数で，$c_{BIC} \left(m \right)$はモデルmに対応するBICによって計算されるウエイトである。詳細は前述Smoothed BICに関する説明を参照のこと。結果はTable1にまとめた。

Table1. 予測の結果

	GARCH	OGARCH	EGARCH	Best-BIC-I	Best-BIC-II	2モデル	3モデル
$MSPE$	95.91	96.26	100.05	96.26	98.77	95.83	97.87

[2]データの出所：Morningstar, Inc..

Best-BIC-Iはローリング予測の中で，GARCH(1,1)とOGARCH(1,1)の中のBICが小さい方で予測した場合の結果を表す。Best-BIC-IIは三つのモデルの場合である。そして，2モデルと3モデルは二つのモデルと三つのモデルによるモデル平均をあらわす。単独のGARCHやOGRCHと比べて，単独のEGARCHの$MSPE$が一番大きい。そして，二つのモデルによるモデル平均の結果はBest-BIC-Iよりも良く一番いい結果を出している。三つのモデル平均の結果はEGARCHモデルのパフォーマンスに影響され悪い結果となっているが，それでもBest-BIC-IIよりよくなっている。

以上の結果は，モデル平均は単純なBICによるモデル選択よりもある程度優れていることを示しているが，実証研究を行う際，必ずしもいつも言えることではない。以上の例に関しても，ローリング・ウインドの幅を変えたり，予測期間を長くしたり，候補となるモデル群を変えたりすることで，異なる結果が出ることになる。モデル平均の長所は理論上のもので，それを実証研究に反映させるために，モデル平均以外の更なる工夫が必要となる。

5　結論

本稿ではモデル選択の拡張であるモデル平均に関して最近の研究成果を概観した。さらに，その未解決問題について論じて，応用例を挙げた。モデル平均はリスクの面で単純なモデル選択より優れている。モデル平均の漸近分布が導出され，モデル平均による推定量のリスクの評価の方法も確立されている。しかし，漸近分布が有限標本分布の良い近似とならず，統計的推論は困難である。この問題を解決すべく本稿は一つの方向を論文の中で述べた。この方向を含め，様々な試みが期待される。

参考文献

Akaike, H. [1973] “Information theory and an extension of the maximum likelihood principle,” in N., P. B. and C. F. eds. *Proc. of the 2nd Int. Symp. on Information Theory*, pp. 267–281.

Bekker, P. A. [1994] “Alternative Approximations to the Distributions of Instrumental Variable Estimators,” *Econometrica*, Vol. 62, No. 3, pp. 657-81, May.

Beran, R. [2000] "REACT Scatterplot Smoothers: Superefficiency through Basis Economy," *Journal of the American Statistical Association*, Vol. 95, No. 449, pp. 155-171.

Beran, R. and L. Dumbgen [1998] "Modulation of estimators and confidence sets," *Annals of Statistics*, Vol. 26, No. 5, pp. 1826-1856.

Bollerslev, T. [1986] "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, Vol. 31, No. 3, pp. 307-327, April.

Brownlees, C. T. and G. M. Gallo [2008] "On Variable Selection for Volatility Forecasting: The Role of Focused Selection Criteria," *Journal of Financial Econometrics*, Vol. 6, No. 4, pp. 513-539.

Buckland, S. T., C. Burnham, K. P. Burnham, and N. H. Augustin [1997] "Model selection: an integral part of inference," *Biometrics*, Vol. 53, pp. 603–618.

Claeskens, G. and N. L. Hjort [2008] *Model Selection and Model Averaging*: Cambridge University Press.

Clyde, M. and E. I. George [2004] "Model Uncertainty," *Statistical Science*, Vol. 19, No. 1, pp. 81-94.

Donald, S. G. and W. K. Newey [2001] "Choosing the Number of Instruments," *Econometrica*, Vol. 69, No. 5, pp. 1161-1191, September.

Draper, D. [1995] "Assessment and Propagation of Model Uncertainty," *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 1, pp. 45-97.

Engle, R. F. [1982] "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, Vol. 50, No. 4, pp. 987-1007, July.

Hansen, B. E. [2007] "Least Squares Model Averaging," *Econometrica*, Vol. 75, No. 4, pp. 1175-1189, 07.

Hjort, N. and G. Claeskens [2003] "Frequentist Model Average Estimators," *Journal of the American Statistical Association*, Vol. 98, pp. 879-899, January.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky [1999] "Bayesian model averaging: a tutorial," *Statistical Science*, Vol. 14, No. 4, pp. 382-417. with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors.

Kuersteiner, G. and R. Okui [2009] "Instrument selection by first stage prediction averaging." Unpublished manuscript.

Kullback, S. and R. A. Leibler [1951] "On Information and Sufficiency," *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79–86.

Leeb, H. and B. M. Potscher [2008] "Can One Estimate The Unconditional Distribution of Post-Model-Selection Estimators?" *Econometric Theory*, Vol. 24, No. 3, pp. 338-376.

Liu, Q. and K. Morimune [2006] "A Modified GARCH Model with Spells of Shocks," *Asia-Pacific Financial Markets*, Vol. 12, No. 1, pp. 29-44.

Mallows, C. L. [1973] "Some comments on Cp," *Technometrics*, Vol. 15, pp. 661–675.

Morimune, K. [1983] "Approximate Distributions of k-Class Estimators When the Degree of Overidentifiability Is Large Compared with the Sample Size," *Econometrica*, Vol. 51, No. 3, pp. 821-841, May.

Nelson, D. B. [1991] "Conditional Heteroskedasticity in Asset Returns: A New Approach," *Econometrica*, Vol. 59, No. 2, pp. 347-70, March.

Pesaran, M. H., C. Schleicher, and P. Zaffaroni [2009] "Model averaging in risk management with an application to futures markets," *Journal of Empirical Finance*, Vol. 16, No. 2, pp. 280–305.

Potscher, B. M. [2006] "The Distribution of Model Averaging Estimators and an Impossibility Result Regarding Its Estimation," MPRA Paper 73, University Library of Munich, Germany.

Schwarz, G. [1978] "Estimating the Dimension of a Model," *The Annals of Statistics*, Vol. 6, No. 2, pp. 461–464.

Stone, M. [1977] "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion," *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39, No. 1, pp. 44–47.

小西貞則・北川源四郎 [2004] 『情報量基準』, 朝倉書店.

竹内啓 [1976] 「情報統計量の分布とモデルの適切さの基準」,『数理科学』, 第153巻, 12-18頁.

森棟公夫 [1985] 『経済モデルの推定と検定』, 共立出版.