Japan Industrial Management Association

J Jpn Ind Manage Assoc **51**, 59–69, 2000

Original

A Study on Bayes Optimal Prediction for Linear Regression Models

Tomohiko SUZUKI,[1] Masayuki GOTOH [2] and Nobuhiko TAWARA[3]

Abstract

In this paper, we propose an asymptotic Bayes optimal prediction algorithm for linear regression model, which reduces complexity in terms of calculation. In the field of industrial engineering, linear regression analyses are mainly applied to statistical quality control and demand forecasting, owing to effectiveness of control, prediction, analysis of structure, and so on. Recently, statistical model selection has been studied as a method of estimation for linear regression models, and applied to various problems of prediction. The statistical model selection is to select a particular model out of all candidates which include the true probabilistic model. The conventional criteria for model selection are F-value, FPE and information criteria; for example, AIC, BIC, and MDL. The mainly purposes of statistical model selection are to detect the true probability model, predict for future observations, and compress the data. Since statistical model selection has many applications, it has been studied not only in the field of statistics but also in various fields of science such as information theory, automatical control theory, and so on. In the case of estimation of the linear regression model by statistical model selection, generally, a particular model is selected by information criterion based on a previous observation from all candidates. In the linear regression analysis, the model class is a set of the combination of explanatory variables. However, in the case of prediction, it is not necessary to select a particular model. In this case, the purpose of prediction is to acquire the accuracy estimator of the future observation. Therefore, previous studies using statistical model selection for prediction may be insufficient. On the other hand, the prediction method based on Bayes decision theory has been reported in various fields. In this method, predictions using the mixture model, which is mixing all candidates, are acquired as Bayes optimal solution, which minimizes the Bayesian mean square error. For this reason, we apply the mixture model for the linear regression models for prediction. We, at first, show that prediction by the mixture model is Bayes optimal prediction. However, it is difficult to strictly calculate the mixture probability because of the integration complexity on the parameter space. Therefore, we propose a new prediction method which removes the integration on account of reducing the complexity. Strictly speaking, we propose an asymptotic Bayes optimal prediction, which calculates the asymptotic posterior predictive distribution; i.e., mixture model. At last, we verify the effectiveness of the proposal through the simulation experiments.

Key words: linear regression model, statistical model selection, bayes decision theory, mixture model

[1]Canon Inc.
[2]Waseda University
[3]Musashi Institute of Technology
Received: April 10, 1998
Accepted: November 25, 1999

NII-Electronic Library Service

Japan Industrial Management Association

論　　文

線形回帰モデルのベイズ最適な予測法に関する研究

鈴木友彦[1]，後藤正幸[2]，俵　信彦[3]

本報では，線形回帰モデルを用いた予測問題において，計算量を抑えた近似ベイズ最適な予測アルゴリズムを提案する．近年，線形回帰モデルの推定方法として，統計的モデル選択が幅広く研究され，予測問題に応用されている．しかし，選択されたモデルに予測値が大きく依存することから，モデルによっては，予測の精確さが悪くなるという問題を含んでいる．そこで本報ではまず，線形回帰モデルを用いた予測問題に限定した場合，ベイズ決定理論を導入し，複数のモデルの混合モデルを用いて予測を行うことが，平均損失最小となることを示す．しかし，混合モデルの計算は複雑となるため，漸近近似的に計算量を抑えた予測法の提案を行う．シミュレーション実験の結果から，提案予測法は，精度・真度ともに，従来法と比較して，より有効であることが明らかとなった．

キーワード：回帰分析，統計的モデル選択，ベイズ決定理論，混合モデル

1. はじめに

線形回帰分析 (Linear regression analysis) の応用領域は，自然科学，工学，人文・社会科学に至るまで，非常に広範囲に及んでいる．とくに経営工学分野においては，統計的品質管理 (SQC) や需要予測などの分野で，制御，予測，構造解析，変動要因解析等に有効な手法として，広く活用されている．

近年，線形回帰モデルの推定方法として，統計的モデル選択が幅広く研究されている．統計的モデル選択は，複数の確率モデルの候補の中から，一つのモデルを選択するという問題で，おもに，真のモデルの推定や予測，情報圧縮等の目的で利用されている．また，その応用範囲が多岐にわたることから，統計学の分野だけでなく，情報理論の分野など様々な学問分野で幅広く研究がなされている．この統計的モデル選択では，モデルを選択する規準として一般に，F 値や FPE，情報量規準 AIC，BIC，MDL 等が用いられている．

線形回帰モデルを用いた予測問題に，統計的モデル選択を応用する場合，モデル構造をサンプルデータ (学習データ) に基づいて一つに特定し，このモデルを用いて予測を行うことになる．しかし，回帰モデルの導出目的を未知のデータ (未学習データ) に対する予測問題のみに限定した場合においては，このとき求められるのは，より精確な予測値を得ることである．したがって，精確な予測値を得ることと，一つのモデルを用いて予測を行うことは直接結びつかず，予測問題に対しては従来法が必ずしも妥当ではないと考えることができる．これに対して，ベイズ決定理論の分野では，予測の際にモデルを一つに限定せずに，考え得る全てのモデルの重み付け平均をとり，このモデル (混合モデル：Mixture) を用いて予測を行うことが平均損失の面で最適となることが知られている〔2〕,〔3〕,〔5〕．

本報ではまず，線形回帰モデルの予測問題においても，全てのモデル候補について事後確率による重み付け平均をとり，この混合モデルを用いて予測を行うことが平均損失最小，すなわちベイズ最適な予測法であることを示す．しかし，任意の事前分布に対して厳密に混合をとる計算は，パラメータ空間上の複雑な積分操作が必要となり，計算量的な困難を生じる．そこで，従来ニューラルネット (NN) モデルに適用された方法〔3〕を線形回帰モデルに適用し，漸近展開を用いて積分操作を排除した簡便な予測法を提案する．NN モデルは Fisher 情報行列が退化する非線形モデルであるので，厳密にはこの漸近式が成り立たないのに対し，本稿のモデルではこの予測法は厳密に漸近ベイズ最適となる．最後に，シミュレーション実験により，提案法の有効性の検証を行う．

2. 問題設定

2.1 線形回帰モデルを用いた予測問題

説明変数が p 個，データが n 組ある場合を考える．説明変数のベクトル列 $x^n = (x_1, \cdots, x_n)$，$(x_i = (x_{i1}, \cdots, x_{ip}) \in \mathcal{R}^p, (i = 1, \cdots, n))$，($\mathcal{R}^p$ は p 次元ユークリッド空間) に対する目的変数のベクトル列を $y^n = (y_1, \cdots, y_n)$ とする．

以上の条件のもとで，本報における予測問題を，次のように定義する．

定義 1 (予測問題) n 組のデータ $(x^n, y^n) = (x_1, y_1), \cdots, (x_n, y_n)$ に基づいて，x_{n+1}が与えられたときの

[1]キャノン株式会社
[2]早稲田大学
[3]武蔵工業大学
受付：1998 年 4 月 10 日，再受付（4 回）
受理：1999 年 11 月 25 日

NII-Electronic Library Service

Japan Industrial Management Association

y_{n+1} の推定値 $\hat{y}_{n+1}$ を得ることを y_{n+1} の予測とする．また，予測の評価規準は，真値 y_{n+1} と推定値 $\hat{y}_{n+1}$ の二乗誤差とする． □

2.2 対象問題の定式化

$g\left(x_i, \theta^{k_m}, m\right)$ を回帰モデルの出力，θ^{k_m} をモデル m の k_m 次元パラメータとすると，モデル m による y_i の確率構造は，

$$y_i = g\left(x_i, \theta^{k_m}, m\right) + \varepsilon_i \qquad (1)$$

となる．ただし，便宜的に $g\left(x_i, \theta^{k_m}, m\right)$ と記述するが，パラメータには i.i.d. のノイズ ε_i の分散も要素として含まれているものとする (ノイズの平均値は 0)．(1) 式において θ^{k_m}，m が定まれば，y^n の同時確率密度関数は (2) 式のように定められる．

$$p\left(y^n \middle| x^n, \theta^{k_m}, m\right) = \prod_{i=1}^{n} p\left(y_i \middle| x_i, \theta^{k_m}, m\right) \qquad (2)$$

また，説明変数 x_i が与えられたときの目的変数 y_i の確率分布が平均 $g\left(x_i, \theta^{k_m}, m\right)$，分散 σ^2 の正規分布にしたがうと仮定すると，モデルの尤度は，

$$\begin{aligned} & p\left(y^n \middle| x^n, \theta^{k_m}, m\right) \\ & = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \cdot \right. \\ & \qquad \left. \sum_{i=1}^{n} \left(y_i - g\left(x_i, \theta^{k_m}, m\right)\right)^2 \right\} \end{aligned} \qquad (3)$$

で与えられる．

3. 従 来 の 研 究

3.1 回帰分析におけるモデル選択

回帰分析において，重要な問題の一つに変数選択がある．これは目的変数 y を予測 (説明) するために，これと関係すると推測される説明変数 x_i をどのように選択し，回帰式に取り込むかという問題である．説明変数が足りない場合には，予測値に偏りを生じ，余分な場合には，推定精度が低下し，予測値がばらつくといった問題や，説明変数間の多重共線性の問題が生じるため，重要な問題として幅広く研究されている．

この変数選択問題は，統計的モデル選択の枠組みの中で解決することができる．統計的モデル選択は，複数の確率モデルの候補の中から，一つのモデルを選択するという問題で，モデルを選択する規準として一般に，F 値や FPE，情報量規準 AIC，BIC，MDL 等が用いられている〔1〕．

統計的モデル選択を回帰モデルの導出に応用する場合，問題は，どの説明変数を取り入れたモデルを選択するかということになる．例えば，説明変数として x_1 から x_{10} まで 10 個の変数が考えられる場合，考え得るモデル候補は，

- x_1 のみ取り入れるモデル
- x_1 と x_2 を取り入れるモデル
- x_1，x_2 と x_3 を取り入れるモデル

⋮

- x_1 から x_{10} まですべての変数を取り入れるモデル

というように，変数の組合わせ数のモデル，この場合は $2^{10} = 1024$ 個のモデルとなるが，これらのモデル候補の中から，情報量規準等を適用することにより，一つのモデルを選択するという問題になる．

3.2 従来研究の問題点

回帰モデルを導出する際に，情報量規準を用いる場合，どの規準を用いる場合においても，一つのモデルを選択し，解析を行うことになる．これは，モデルの導出目的が，真のモデルを発見するという目的では妥当であるが，未知のデータに対する予測という目的においては，他のモデルの可能性を無視していることから，精度の高い予測値を得られなくなる危険性も含んでいる．

予測問題において最も重要なことは，精確な予測値を得ることであり，従来法のように，一つのモデルを用いて予測を行わなくとも，精確な予測値が得られればよいことになる．そこで本報では，予測問題に対し，モデルを 1 つ選択するという制約を設けず，予測誤差の最小化を考える．

4. ベイズ決定理論に基づく予測式の導出

4.1 ベイズ最適な予測法の導出

本報では，二乗誤差損失のもとで，ベイズ規準を導入し，候補となるすべてのモデルの事後確率による重み付け平均モデルを用いて予測を行う方法を検討する．

モデル族 $\mathcal{M}$ に含まれる線形回帰モデルを m，θ^{k_m} は m によって定められる k_m 次元パラメータとし，候補となるモデルが T 個ある場合を考える．また，決定関数を $Dy\left(x_{n+1}, x^n, y^n\right)$ とし，損失関数として二乗誤差 $L\left(y_{n+1}, Dy\left(x_{n+1}, x^n, y^n\right)\right)$ を用いる．リスク関数 $R(\theta^{k_m}, m)$ は，決定関数 $Dy\left(x_{n+1}, x^n, y^n\right)$ を用いたときのモデルの期待損失で定義される．損失関数，リスク関数はそれぞれ，

$$\begin{aligned} & L\left(y_{n+1}, Dy\left(x_{n+1}, x^n, y^n\right)\right) \\ & \qquad = \left(y_{n+1} - Dy\left(x_{n+1}, x^n, y^n\right)\right)^2 \end{aligned} \qquad (4)$$

NII-Electronic Library Service

Japan Industrial Management Association

$$R(\theta^{k_m},m)=\int_{y_{n+1}} L\left(y_{n+1},Dy\left(x_{n+1},x^n,y^n\right)\right) p\left(y_{n+1}|x_{n+1},\theta^{k_m},m\right)dy_{n+1} \quad (5)$$

となる．ここで，リスク関数を最小にする決定関数を定めればよいことになるが，任意の θ^{k_m},m に対して，(5) 式を最小化する決定関数は存在しない．そこで，ベイズ決定理論では，リスク関数を事後確率 $p\left(\theta^{k_m},m|x^n,y^n\right)$ で平均化したベイズリスクの最小化を行う．

$$\begin{aligned}BR&=\int_{y_{n+1}} L\left(y_{n+1},Dy\left(x_{n+1},x^n,y^n\right)\right)\\&\quad\sum_m\int_{\theta^{k_m}} p\left(y_{n+1}|x_{n+1},\theta^{k_m},m\right)\\&\qquad p\left(\theta^{k_m},m|x^n,y^n\right)d\theta^{k_m}dy_{n+1}\\&=\int_{y_{n+1}} L\left(y_{n+1},Dy\left(x_{n+1},x^n,y^n\right)\right)\\&\qquad p_{mix}\left(y_{n+1}|x_{n+1},x^n,y^n\right)dy_{n+1}\end{aligned} \quad (6)$$

となる．ここで，$p_{mix}\left(y_{n+1}|x_{n+1},x^n,y^n\right)$ は事後混合分布と呼ばれ，

$$\begin{aligned}&p_{mix}\left(y_{n+1}\middle|x_{n+1},x^n,y^n\right)\\&=\sum_m\int_{\theta^{k_m}} p\left(y_{n+1}\middle|x_{n+1},\theta^{k_m},m\right)\\&\qquad p\left(\theta^{k_m},m|x^n,y^n\right)d\theta^{k_m}\end{aligned} \quad (7)$$

である．(6) 式のベイズリスクを最小にする決定関数は，補題 1 で与えられる〔5〕．

補題 1 ベイズリスク BR を最小にする決定関数 $Dy\left(x_{n+1},x^n,y^n\right)$ は，

$$\begin{aligned}&Dy\left(x_{n+1},x^n,y^n\right)\\&=\int_{y_{n+1}} y_{n+1}p_{mix}\left(y_{n+1}|x_{n+1},x^n,y^n\right)dy_{n+1}\end{aligned} \quad (8)$$

で与えられる．□

以上から，二乗誤差損失を最小にする，すなわちベイズ最適な予測を行うには事後混合分布 $p_{mix}\left(y_{n+1}|x_{n+1},x^n,y^n\right)$ の期待値を用いればよいことが分かる．

4.2 漸近近似を用いた予測法

ベイズ最適な予測を行うには，事後混合分布の期待値を用いればよいことを示したが，一般的なモデル族と事前密度に対しての事後混合分布の厳密な計算は，パラメータ空間上の複雑な積分操作が必要になり，計算量的な困難を生じる．そこで，線形回帰モデル上で成り立つ事後確率の漸近正規性を直接用いることによりこの積分操作を排除し，各モデルの事後確率を漸近評価することにより，漸近近似的に事後混合分布を求める方法を提案する〔3〕,〔4〕．この方法は，Laplace の漸近展開を用いて同様の近似式を導いた〔3〕の方法と同じ結果になるが，このことは線形モデルでは本質的と考えられる．

定理 1 事前密度 $f(\theta^{k_m}|m)$ は連続微分可能であり，$\forall\theta^{k_m}$ に対して $f(\theta^{k_m}|m)>0$ であるとする．このとき，線形回帰モデルに対する (7) 式の事後混合分布は漸近的に，

$$\begin{aligned}&p_{mix}\left(y_{n+1}|x_{n+1},x^n,y^n\right)\\&=\sum_m\lambda_{k_m}p\left(y_{n+1}\middle|x_{n+1},\hat\theta^{k_m},m\right)+o(1)\end{aligned} \quad (9)$$

となる．ここで，λ_{k_m} は Z を規準化定数とし，

$$\lambda_{k_m}=\left[p\left(y^n\middle|x^n,\hat\theta^{k_m},m\right)f\left(\hat\theta^{k_m}\middle|m\right)p(m)\left(\frac{n}{2\pi}\right)^{-\frac{k_m}{2}}\sqrt{\det I\left(\hat\theta^{k_m}\middle|m\right)}^{-1}\right]\Big/Z \quad (10)$$

$$Z=\sum_m p\left(y^n\middle|x^n,\hat\theta^{k_m},m\right)f\left(\hat\theta^{k_m}\middle|m\right)p(m)\left(\frac{n}{2\pi}\right)^{-\frac{k_m}{2}}\sqrt{\det I\left(\hat\theta^{k_m}\middle|m\right)}^{-1} \quad (11)$$

で与えられる．また，$\hat\theta^{k_m}$ は最尤推定量，$p(m)$ はモデルの事前分布，k_m はモデル m のパラメータ数である．

(証明) パラメータの事前確率 $f\left(\theta^{k_m}\middle|m\right)$ が正で微分可能であれば，パラメータの事後確率 $f\left(\theta^{k_m}\middle|x^n,y^n,m\right)$ は漸近的に正規分布にしたがうことが知られている〔6〕．すなわち，

$$\begin{aligned}&\frac{1}{\sqrt{n}^{k_m}}f\left(\theta^{k_m}\middle|x^n,y^n,m\right)\to\frac{\sqrt{\det I\left(\hat\theta^{k_m}\middle|m\right)}}{\sqrt{2\pi}^{k_m}}\\&\exp\left\{-\frac{1}{2}\left(\theta^{k_m}-\hat\theta^{k_m}\right)^TI\left(\hat\theta^{k_m}\middle|m\right)\left(\theta^{k_m}-\hat\theta^{k_m}\right)\right\}\end{aligned} \quad (12)$$

NII-Electronic Library Service

Japan Industrial Management Association

である．したがって，θ^{k_m} に $\hat{\theta}^{k_m}$ を代入すると，

$$\frac{1}{\sqrt{n}^{k_m}} f\left(\hat{\theta}^{k_m}\middle|x^n,y^n,m\right) \rightarrow \frac{\sqrt{\det I\left(\hat{\theta}^{k_m}\middle|m\right)}}{\sqrt{2\pi}^{k_m}} \tag{13}$$

となる．また，ベイズの定理から，

$$p\left(x^n,y^n|m\right)=\frac{p\left(y^n\middle|x^n,\hat{\theta}^{k_m},m\right)f\left(\hat{\theta}^{k_m}\middle|m\right)}{f\left(\hat{\theta}^{k_m}\middle|x^n,y^n,m\right)} \tag{14}$$

であるから，(13) 式を (14) 式に代入することにより，

$$p\left(x^n,y^n|m\right) \approx \frac{\sqrt{2\pi}^{k_m} p\left(y^n\middle|x^n,\hat{\theta}^{k_m},m\right)f\left(\hat{\theta}^{k_m}\middle|m\right)}{\sqrt{n}^{k_m}\sqrt{\det I\left(\hat{\theta}^{k_m}\middle|m\right)}} \tag{15}$$

が得られる．さらに，ベイズの定理により，

$$P\left(m|x^n,y^n\right) \propto p\left(x^n,y^n|m\right)P(m) \tag{16}$$

となるので，(15) 式から，

$$P\left(m|x^n,y^n\right) \propto \frac{\sqrt{2\pi}^{k_m} p\left(y^n\middle|x^n,\hat{\theta}^{k_m},m\right)f\left(\hat{\theta}^{k_m}\middle|m\right)P(m)}{\sqrt{n}^{k_m}\sqrt{\det I\left(\hat{\theta}^{k_m}\middle|m\right)}} \tag{17}$$

となる．一方，$p\left(y_{n+1}\middle|x_{n+1},\theta^{k_m},m\right)$ は θ^{k_m} に対し連続微分可能，かつ $p\left(y_{n+1}\middle|x_{n+1},\theta^{k_m},m\right)<C$ となる C が存在する．また，漸近正規性から，

$$\int_{B_\delta} f\left(\theta^{k_m}\middle|x^n,y^n,m\right)d\theta^{k_m} \rightarrow 1 \tag{18}$$

$$\int_{\overline{B_\delta}} f\left(\theta^{k_m}\middle|x^n,y^n,m\right)d\theta^{k_m} \rightarrow 0 \tag{19}$$

ただし，

$$B_\delta=\left\{\theta^{k_m}\middle|\parallel\theta^{k_m}-\hat{\theta}^{k_m}\parallel<\delta\right\} \tag{20}$$

となる．これより，

$$\begin{aligned}&\int_{\theta^{k_m}} p\left(y_{n+1}\middle|x_{n+1},\theta^{k_m},m\right) f\left(\theta^{k_m}\middle|x^n,y^n,m\right)d\theta^{k_m}\\ &=\int_{B_\delta} p\left(y_{n+1}\middle|x_{n+1},\theta^{k_m},m\right) f\left(\theta^{k_m}\middle|x^n,y^n,m\right)d\theta^{k_m}\\ &\quad+\int_{\overline{B_\delta}} p\left(y_{n+1}\middle|x_{n+1},\theta^{k_m},m\right) f\left(\theta^{k_m}\middle|x^n,y^n,m\right)d\theta^{k_m}\\ &\rightarrow\int_{B_\delta} p\left(y_{n+1}\middle|x_{n+1},\theta^{k_m},m\right) f\left(\theta^{k_m}\middle|x^n,y^n,m\right)d\theta^{k_m}\end{aligned} \tag{21}$$

が得られる．ここで，$\theta^{k_m}\in B_\delta$ に対しては，$\delta\rightarrow+0$ のとき，

$$p\left(y_{n+1}\middle|x_{n+1},\theta^{k_m},m\right) \rightarrow p\left(y_{n+1}\middle|x_{n+1},\hat{\theta}^{k_m},m\right) \tag{22}$$

から，

$$\int_{\theta^{k_m}} p\left(y_{n+1}\middle|x_{n+1},\theta^{k_m},m\right) f\left(\theta^{k_m}\middle|x^n,y^n,m\right)d\theta^{k_m} \rightarrow p\left(y_{n+1}\middle|x_{n+1},\hat{\theta}^{k_m},m\right) \tag{23}$$

が得られる．ここで，(7) 式の事後混合分布は，

$$\begin{aligned}&p_{mix}\left(y_{n+1}\middle|x_{n+1},x^n,y^n\right)\\ &=\sum_m p\left(m|x^n,y^n\right)\int_{\theta^{k_m}} p\left(y_{n+1}|x_{n+1},\theta^{k_m},m\right)\cdot f\left(\theta^{k_m}|x^n,y^n\right)d\theta^{k_m}\end{aligned} \tag{24}$$

と変形できるので，(24) 式に (17)，(23) 式を代入することにより，

$$p_{mix}\left(y_{n+1}|x_{n+1},x^n,y^n\right) \propto \frac{\sqrt{2\pi}^{k_m} p\left(y^n\middle|x^n,\hat{\theta}^{k_m},m\right)f\left(\hat{\theta}^{k_m}\middle|m\right)p(m)}{\sqrt{n}^{k_m}\sqrt{\det I\left(\hat{\theta}^{k_m}\middle|m\right)}} \cdot p\left(y_{n+1}\middle|x_{n+1},\hat{\theta}^{k_m},m\right) \tag{25}$$

が得られる．□

〔3〕において提案された NN モデルに対する予測法は，厳密には Laplace 展開が適用できない NN モデルに対し形式的に適用したものであるのに対し，定理 1 は正しい漸近式となっており，従って漸近的にベイズ最適を保証するという理論的根拠を持つ．また本稿では，線形回帰モデルが持つ事後確率密度の漸近正規性を直接用いることにより，直感的にも理解し易い

NII-Electronic Library Service

Japan Industrial Management Association

証明を与えている．すなわち，モデルの事後確率の収束速度は結局パラメータの推定精度に依存することになり，パラメータの事後密度が正規分布となることから Fisher 情報行列や $\sqrt{2\pi}^{k_m}$ の項が出てくることがわかる．

また，本稿では線形回帰モデルを取り上げているが，定理1の証明では事後密度の漸近正規性が本質的であり，これが成り立つモデル族に対しては定理1が一般的に成り立つことがわかる．このことは今後の拡張性の面で重要である．漸近正規性が成り立つモデル族については〔6〕などを参照．

(7) 式の θ^{k_m} が，(9) 式において，その最尤推定量 $\hat{\theta}^{k_m}$ で置き換わっていることに注意すると，事後混合分布の計算は，積分操作を行わなくても，最尤推定量を代入することによって，漸近近似的に和の形で与えられることを示している．本報では回帰係数の推定方法として，最小二乗法を用いるが，この推定量が最尤推定量であることから，事後混合分布の計算が，(9) 式により可能となる．また λ_{k_m} が，漸近的にモデルの事後確率を表していることが，(7) 式と (9) 式から分かる．

したがって，(3) 式および (10) 式から，

$$\lambda_{k_m} = \left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - g\left(x_i, \hat{\theta}^{k_m}, m\right) \right)^2 \right\} f\left(\hat{\theta}^{k_m} \middle| m\right) p(m) \sqrt{\det I\left(\hat{\theta}^{k_m} \middle| m\right)}^{-1} \left(\frac{n}{2\pi} \right)^{-\frac{k_m}{2}} \right] \Big/ Z \qquad (26)$$

とおくと，線形回帰モデルの事後混合分布の漸近式が得られる．

また，ベイズリスクを漸近近似的に最小にする決定関数 $Dy\left(x_{n+1}, x^n, y^n\right)$ は，定理1より (8) 式と (9), (26) 式用いて次のように求めることができる．

定理 2 ベイズリスクを漸近近似的に最小にする決定関数 $Dy\left(x_{n+1}, x^n, y^n\right)$ は，

$$Dy\left(x_{n+1}, x^n, y^n\right) = \sum_m \lambda_{k_m} g\left(x_{n+1}, \hat{\theta}^{k_m}, m\right) \qquad (27)$$

で与えられる．ここで，$g\left(x_{n+1}, \hat{\theta}^{k_m}, m\right)$ は $\hat{\theta}^{k_m}$ を用いた線形回帰モデル m にデータ x_{n+1} を入力した際の出力である． □

以上から，ベイズ最適な予測を行うには，回帰モデルの出力 $g\left(x_{n+1}, \hat{\theta}^{k_m}, m\right)$ の事後確率による重み付け和を用いればよいことが分かる．

5. シミュレーション実験

5.1 評価規準

提案法の有効性を検証するために，比較対象に AIC による予測を取り上げ，シミュレーション実験を行う．評価規準には，予測精度および真度を取り上げる．そこで，テスト用の未学習データに対する提案法および AIC の複数の予測値と真値の二乗誤差の平均，分散で評価を行うことにする．

5.2 シミュレーション条件

説明変数を 5 個含む重回帰モデルを用いた．このとき，候補となるモデル数は 32 個である．また，モデルの事前分布には一様分布，パラメータの事前分布には，次式で表わされる Jeffreys' prior を用いた．

$$f\left(\hat{\theta}^{k_m} \middle| m\right) = \frac{\sqrt{\det I\left(\hat{\theta}^{k_m} \middle| m\right)}}{\int \sqrt{\det I\left(\theta^{k_m} \middle| m\right)} d\theta^{k_m}} \qquad (28)$$

実験は，真のパラメータ (回帰係数) の設定を以下のように 3 パターンについて行った．学習データの各説明変数は独立に標準正規分布 $N(0,1)$ で発生させた．また，学習データ数は 10〜100 まで 5 間隔，テスト用の未学習データ数にはノイズを加えないものを 1000 個用い，シミュレーション回数は 1,000 回とした．

実験 1
真のパラメータ数 ： 3
真のパラメータ β^* ：
$(\beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*, \beta_5^*) = (0.1, 0, 0.2, 0, 10.0)$
特徴 ： 3 つの変数が効いているが，その中でも 1 つの変数の影響がとくに大きい．

実験 2
真のパラメータ数 ： 3
真のパラメータ β^* ：
$(\beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*, \beta_5^*) = (0.1, 0, 5.0, 0, 10.0)$
特徴 ： 3 つの変数が効いているが，その中でも 1 つの変数の影響がとくに小さい．

実験 3
真のパラメータ数 ： 3
真のパラメータ β^* ：
$(\beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*, \beta_5^*) = (1.0, 0, 3.0, 0, 5.0)$
特徴 ： 取り上げた変数の中で，3 つの変数が同程度の割合で効いている．

NII-Electronic Library Service

Japan Industrial Management Association

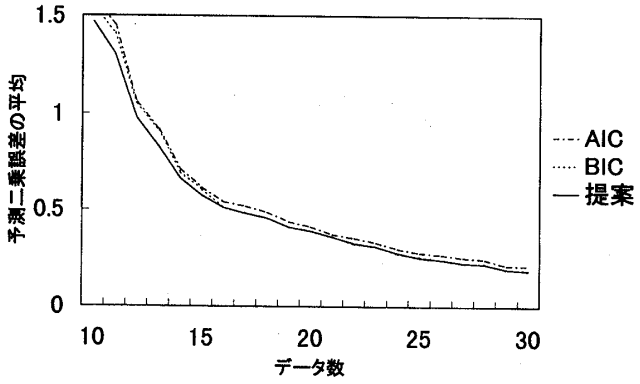

図1 実験 3 (予測時の説明変数分散 1, 誤差分散 1) の予測誤差の平均

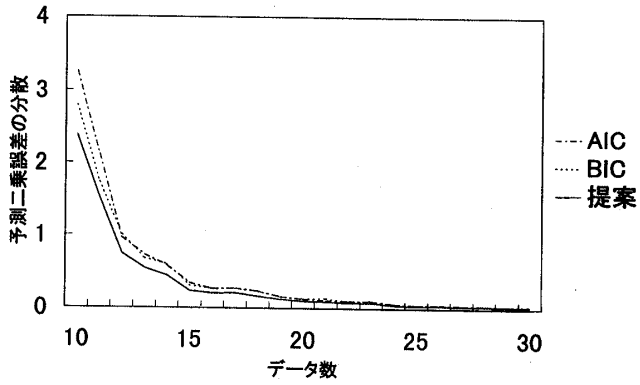

図2 実験 3 (予測時の説明変数分散 1, 誤差分散 1) の予測誤差の分散

さらに，学習時 (解析時) と予測時の説明変数の変動の差による影響，及び誤差分散の大きさによる影響を調べるため，実験 1 ～ 3 においてそれぞれ，

(1) 予測時の説明変数の分散が 1 の場合と 2 の場合 (学習時はいずれも 1)

(2) 誤差分散が 1 の場合と 2 の場合

を実験した．誤差分散が 2 のケースは，説明変数の変動に比べて誤差項の影響が大きい状況である．

5.3 結果および考察

3 パターンの実験の結果，ほぼ同様の結果が得られたので，ここでは，実験 3 の結果を拡大 (学習データ数：10 から 30 まで 1 間隔，シミュレーション回数：10,000 回) して考察を行う．以下に，実験 3 の予測データに対する平均二乗誤差の実験繰り返し 10000 回による平均値，および分散，および各モデルの事後確率の推移を示す．**図 1,2** は予測時の説明変数の分散が 1，誤差分散も 1 の場合である．**図 3** にはこの実験での事後確率の推移を示しており，横軸は考えうる $2^5 = 32$ 個のモデルを説明変数の少ないモデルから順番に番号付けしたものである．**図 4,5** は先の設定から予測時の説明変数の分散を 2 とした場合であり，**図 6,7** は誤差分散を 2 とした場合である．

実験 1 から実験 3 を通じて，提案法と従来法の違いをまとめると以下のようになる．

（1） 提案法は，予測精度の評価規準である平均二乗誤差の平均，および安定した予測を行えるかの評価規準である平均二乗誤差の分散ともに従来法を下回っている．とくに，二乗誤差の分散でその差が顕著になる．

（2） 分散に関しては，学習データ数がとくに少ないとき，提案法の有効性が顕著になる．

（3） 提案法において，各モデルの事後確率は真のモデルによって異なる．また，真のモデルの事後確率が高くなるが，1 への近づき方は遅く，より説明変数の多いモデルの事後確率も大きな値をとる．

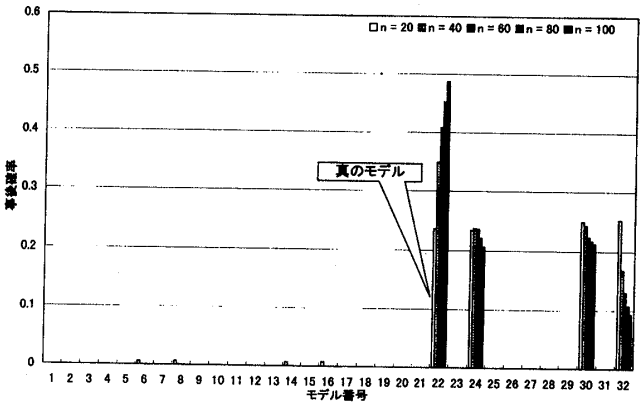

図3 実験 3 の各モデルの事後確率の推移

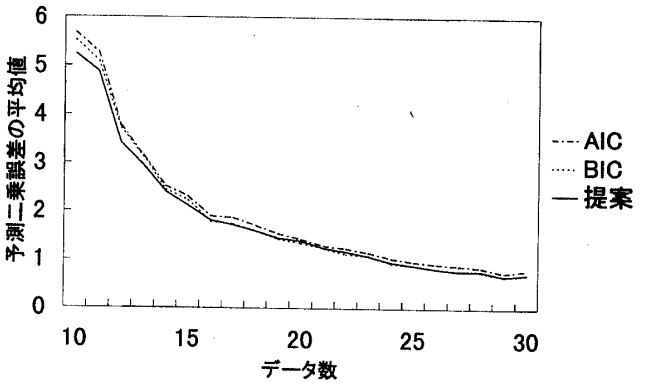

図4 実験 3 (予測時の説明変数分散 2, 誤差分散 1) の予測誤差の平均

まず，実験 10000 回に対する平均二乗誤差の分散が小さいということは，学習データのばらつきに対して予測精度が安定していることを示している．学習データ数が少ないとき，提案法による予測が従来法に比べて安定している理由として，従来法では学習データに

NII-Electronic Library Service

Japan Industrial Management Association

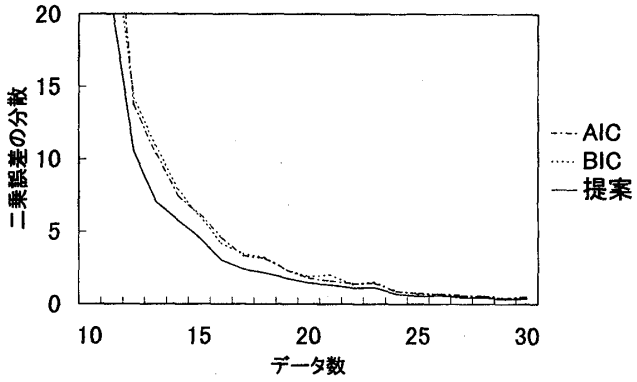

図5 実験 3 (予測時の説明変数分散 2, 誤差分散 1) の予測誤差の平均

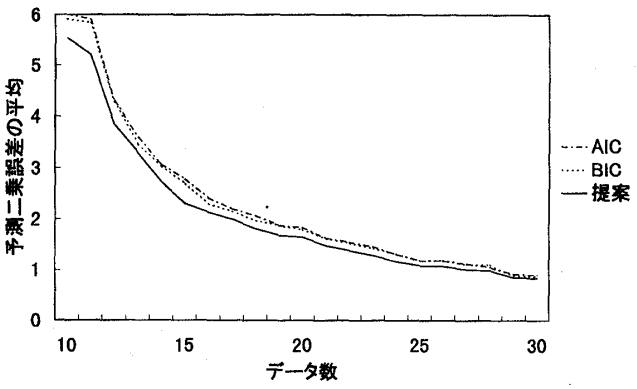

図6 実験 3 (予測時の説明変数分散 1, 誤差分散 2) の予測誤差の平均

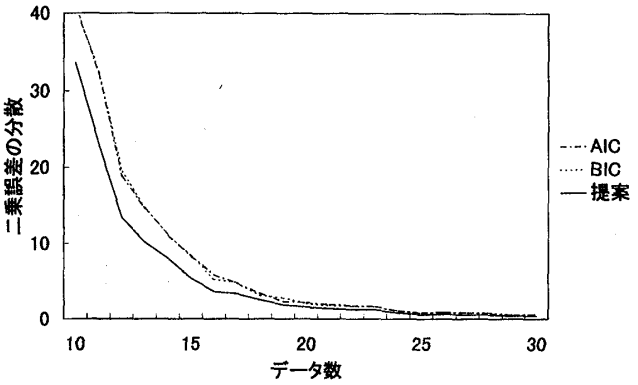

図7 実験 3 (予測時の説明変数分散 1, 誤差分散 2) の予測誤差の平均

オーバーフィットしたモデルから 1 つを選択し予測を行うのに対し，提案法では混合モデルを用いているため，その平滑化の性質が安定した予測を導いていると考えられる．すなわち，“確率的に得られる学習データに対してロバストな予測” という面で提案法の有効性が伺える．

また，(3) については，真のモデルの回帰係数が大きいほど (実験 3)，真のモデルの推定が行われ易く，真のモデルの候補を特定できているのに対し，実験 1 のように影響の小さな変数が多い場合には真のモデルを特定できていないためと考えられる．また，事後確率は真のモデルよりも次数の小さいモデルに対しては速やかに 0 に収束するが，より次数の高いモデルについては値が残っている．これは従来から統計的モデル選択の研究からも明らかにされてきている事実である．

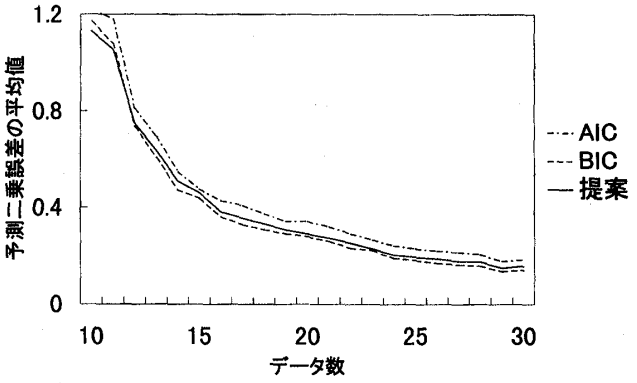

図8 追加実験 A の平均二乗誤差の平均

次に，提案法の特性をさらに深く検討するため，次のような特殊な状況を 2 パターン用意し，実験を行った．ただし，説明変数の分散は 1，誤差分散も 1 である．

追加実験 A

真のパラメータ数 ： 1

真のパラメータ β^* ：

$(\beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*, \beta_5^*) = (0, 0, 0, 0, 5.0)$

特徴 ： 取り上げた変数の中で，実際は 1 つの変数しか効いていない．

追加実験 B

真のパラメータ数 ： 5

真のパラメータ β^* ：

$(\beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*, \beta_5^*) = (1.0, 2.0, 3.0, 4.0, 5.0)$

特徴 ： 取り上げた変数すべてが，ほぼ同程度の割合で効いている．

追加実験の結果を**図 8** から**図 11** に示す．

追加実験の結果から，提案法は真のモデルの次数がモデルクラスの中で最も小さいか，あるいは最も大きいという特殊な状況では，AIC や BIC による予測精度が提案法と同等か，あるいは高くなる場合があることがわかる．これは，AIC が複雑なモデルを選択し易く，逆に BIC はシンプルなモデルを選択し易い性質を持っているため，このような特殊な状況では真のモデルを選択することができるためと考えられる．ただし，この場合でも予測二乗誤差の分散をみると，提案法は

NII-Electronic Library Service

Japan Industrial Management Association

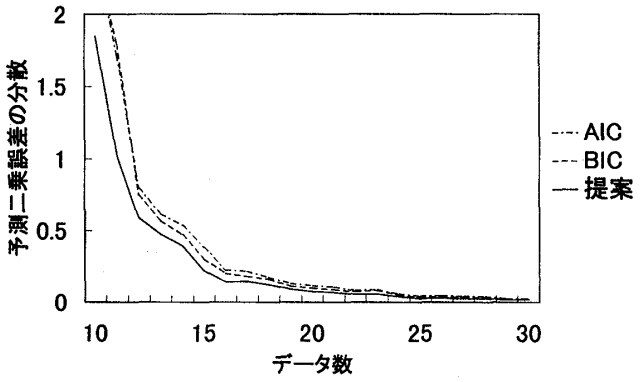

図9　追加実験 A の平均二乗誤差の分散

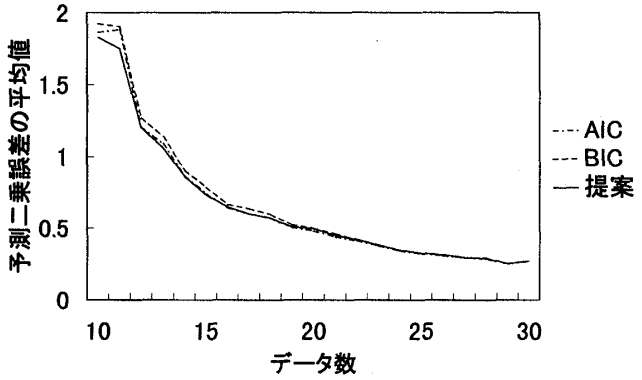

図10　追加実験 B の平均二乗誤差の平均

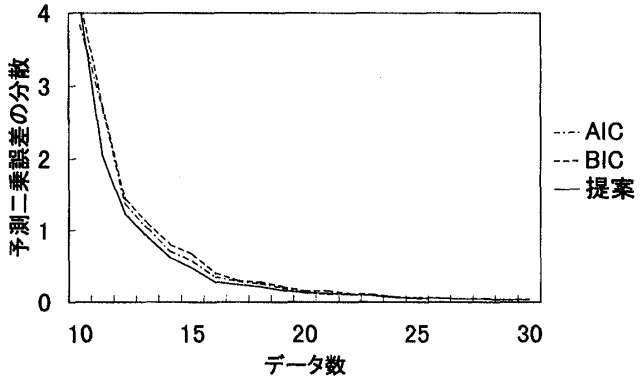

図11　追加実験 B の平均二乗誤差の分散

優れており，ロバストな予測という面からは有効であると考えられる．

以上から，追加実験Bのような状況を除いて，提案法が従来法よりも，平均二乗誤差の平均，分散ともに小さい予測，すなわち精確な予測が行えることが明らかとなった．ただし，今回は無情報事前分布のみを考え，事前確率による差異までは考察していない (事前分布を考慮した場合にはAICなどとの比較には無理が生じる．)．しかし，現実問題では事前分布の意味が対象となる現象と対応付けられる場合が考えられ，このようなより詳細な検討は今後必要と考えられる．

表1　賃貸予測に対する結果（単位：万円^2)

学習データ数	AIC	BIC	提案法
$n=10$	7.83	7.83	7.56
$n=15$	6.17	6.17	5.84
$n=20$	4.56	4.73	4.53

また，今回は説明変数間に相関がある場合，すなわち多重共線性の問題を議論しなかった．しかし，この問題は現実問題で多く起こる本質的な課題であり，その場合のベイズの事後密度の漸近形を調べる必要があり，これは今後の課題とする．

5.4　実際のデータへの適用

本節では，小規模ではあるが，実際のデータ解析へ適用した例を示す．ここでは，予測すべき目的変数 Y を物件の賃貸料とし，説明変数として X_1：最寄り駅からの距離，X_2：物件の広さ，X_3：バス・シャワーの有無，X_4：洋室であるか和室であるか，X_5：築年数とした．これらの物件の条件 $X_1 \sim X_5$ から賃貸料 Y を予測するための回帰モデルを考える．物件は 4.5(万円)～ 15.0(万円) のものを採用した．

需要と供給の関係や経済状況による構造の時間的変動については考えず，構造が定常状態にあるものとする[1]．解析用データとして 都心の主要駅周辺の物件を採用し，評価のためのテストデータも同じ条件で解析用データに含まれないものを 30 件採用した．

以下に，テストデータに対する予測結果を示す．予測誤差は平均二乗誤差である．

$n=10,15$ で AIC と BIC が同じ予測精度を与えているのは，同じモデルを選択したためである．これは，両規準の差異がペナルティ項にあり，それが n が大きいところで差が顕著になるためと考えられる．この結果，若干ではあるが提案法は精度良い予測結果を与えている．ただし，AIC や BIC は回帰モデルを一つ選択しているので，構造解析や予測の信頼区間などの点でより詳細な情報を与えており，その観点からはモデルを選択する利点がある．両手法はどちらかが優れるというものではなく，場面によって両手法を補足的に利用することが自然と考えられる．

また，作為的ではあるが (解析用) 学習データに外れ値と思われる物件を 1 件混入させた場合について示す．

この場合は，差が少し顕著になる．すなわち，提案法がこのような外れ値に対してロバストであることが伺える．モデルを一つ選択する方法では，残差の解析や

[1] 1999 年前半のデータを用いたが，賃貸料に急激な変動は見られないのため構造は安定した状態にあると仮定できる．

NII-Electronic Library Service

Japan Industrial Management Association

表2 賃貸予測に対する結果：外れ値の混入する場合（単位：万円^2）

学習データ数	AIC	BIC	提案法
$n=10$	8.83	8.83	7.26
$n=15$	7.19	7.51	6.02
$n=20$	4.74	4.97	4.62

テコ比やt値を見ることによって外れ値を吟味することができるが，提案法では現在のところそのような手法はない．その反面，提案法は学習データに含まれる外れ値に対してロバストな予測を与える．これは，モデルを一つに限定せず，平均的に用いているためと考えられる．

本来このような外れ値の問題は，理論を適用する実務的問題を深く検討した上で様々な観点から吟味する必要があり，これは今後の課題とする．

6. ま と め

ベイズ決定理論に基づき，候補となる複数のモデル全ての事後混合分布を用いて予測を行うことがベイズ最適であることを示した．また，パラメータ空間での積分操作を排除できる漸近近似式を，パラメータの事後確率密度の漸近正規性を用いて導出し，漸近的にベイズ最適な予測を行う方法を提案した．シミュレーション実験の結果，予測の精度と安定性という面で提案法の有効性を検証することができた．

提案手法は予測という問題のみを扱ったものであり，現在のところ予測の信頼区間などの情報を与えていない．また，構造解析などの目的で回帰分析を行うことができるという点においても，モデルを一つ選択する方法が自然である．したがって，従来の回帰モデルを一つ選択する方法とうまく合わせて，相補的に利用していくべきであると考えられる．

本論文に対し1人の査読者の方から，理論的，および実務的観点による非常に有益なご指摘を戴きました．数回にわたる有益な議論をさせて戴いたことに対し，深く感謝の意を表します．

参 考 文 献

〔1〕 坂元慶行，石黒真木夫，北川源四郎：「情報量統計学」，共立出版 (1983)

〔2〕 韓太舜，小林欣吾：「情報と符号化の数理」，岩波書店 (1994)

〔3〕 橋川弘紀，後藤正幸，俵 信彦："階層型ニューラルネットワークの混合モデルによるベイズ最適な予測について"，電子情報通信学会論文誌 (D-II)，Vol.J80-D-II，No.7，pp.1919–1928 (1997)

〔4〕 B.S. Clarke and A.R. Barron: "Information- Theoretic Asymptotics of Bayes Methods", *IEEE Trans. Information Theory*, Vol.36, pp.453–471 (1990)

〔5〕 T.Matsushima, H.Inazumi, and S.Hirasawa: "A Class of Distortionless Codes Designed by Bayes Decision Theory", *IEEE Trans. Information Theory*, Vol.37, pp.1288–1293 (1991)

〔6〕 J.M. Bernardo and A.F.M. Smith: "Bayesian Theory", John wiley & Sons (1994)

付録 定理2の証明

補題1の(8)式は，(9)，(26)式から，

$$
\begin{aligned}
&Dy\left(x_{n+1},x^n,y^n\right)\\
&=\int_{y_{n+1}} y_{n+1}\quad p_{mix}\left(y_{n+1}|x_{n+1},x^n,y^n\right)dy_{n+1}\\
&\to\sum_m\left[\left\{\int_{y_{n+1}} y_{n+1}p\left(y_{n+1}\middle|x_{n+1},\hat{\theta}^{k_m},m\right)\right.\right.\\
&\left.dy_{n+1}\right\}\left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\right.\right.\\
&\left.\left.\sum_{i=1}^{n}\left(y_i-g\left(x_i,\hat{\theta}^{k_m},m\right)\right)^2\right\}f\left(\hat{\theta}^{k_m}\middle|m\right)p(m)\right.\\
&\left.\sqrt{\det I\left(\hat{\theta}^{k_m}\middle|m\right)}^{-1}\left(\frac{n}{2\pi}\right)^{-\frac{k_m}{2}}\right]\Big/Z\Big] \qquad (29)
\end{aligned}
$$

となる．

ここで，$\int_{y_{n+1}} y_{n+1}p\left(y_{n+1}|x_{n+1},\hat{\theta}^{k_m},m\right)dy_{n+1}$ は $p\left(y_{n+1}|x_{n+1},\hat{\theta}^{k_m},m\right)$ の期待値であり，学習データより導出した線形回帰モデル m に未学習データ x_{n+1} を入力したときの出力であるから，

$$
\begin{aligned}
&\int_{y_{n+1}} y_{n+1}p\left(y_{n+1}\middle|x_{n+1},\hat{\theta}^{k_m},m\right)dy_{n+1}\\
&\qquad=g\left(x_{n+1},\hat{\theta}^{k_m},m\right) \qquad (30)
\end{aligned}
$$

である．

したがって，決定関数 $Dy\left(x_{n+1},x^n,y^n\right)$ は，

$$
\begin{aligned}
&Dy\left(x_{n+1},x^n,y^n\right)\\
&=\sum_m\left[g\left(x_{n+1},\hat{\theta}^{k_m},m\right)\left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\right.\right.\\
&\left.\left.\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-g\left(x_i,\hat{\theta}^{k_m},m\right)\right)^2\right\}\right.\right.
\end{aligned}
$$

NII-Electronic Library Service

Japan Industrial Management Association

$$f\left(\hat{\theta}^{k_m}\middle|m\right)p(m)\sqrt{\det I\left(\hat{\theta}^{k_m}\middle|m\right)}^{-1}\left(\frac{n}{2\pi}\right)^{-\frac{k_m}{2}}\Bigg]\Bigg/Z\Bigg] = \sum_m \lambda_{k_m} g\left(x_{n+1},\hat{\theta}^{k_m},m\right) \tag{31}$$

で与えられる. □

NII-Electronic Library Service