論　文

モデル選択基準 SIC のスパース回帰分析への適用

津田　宏治[†]　　杉山　将[††]　　クラウスロバート　ミュラー[†††]

Subspace Information Criterion for Sparse Regressors

Koji TSUDA[†], Masashi SUGIYAMA[††], and Klaus-Robert MULLER[†††]

あらまし　ℓ_1 正則化項を加えて線形回帰問題を解くと，パラメータベクトルのほとんどの要素が 0 であるような解（スパースな解）が得られる（スパース回帰分析）．スパース回帰分析は特徴選択等に有効であるが，汎化誤差を小さくするには，正則化定数を最適に設定する必要がある．本論文では，スパース回帰分析の正則化定数を設定するための基準として，従来から知られている Subspace Information Criterion を拡張し，Generalized Subspace Information Criterion (GSIC) を提案する．さらに，GSIC にバイアスを導入することによって安定性を増した改良版 (Biased GSIC) についても述べる．

キーワード　モデル選択，Subspace Information Criterion (SIC), スパース回帰分析，カーネル法

1. はじめに

教師つき学習においては，汎化誤差の増大の原因となる過学習を避けるため，正則化が行われることが多い [16]．正則化学習においては，パラメータベクトル $\boldsymbol{\theta}$ は，訓練誤差と正則化項 R の重みつき和

$$\mathrm{Error}(\boldsymbol{\theta}) = \mathrm{TrainingError}(\boldsymbol{\theta}) + \lambda R(\boldsymbol{\theta}) \qquad (1)$$

を最小にするように決定される．ここで λ は正則化定数と呼ばれる．従来，最もよく使われる正則化項は，パラメータの二次関数である．例えば，Weight decay [16] においては，パラメータベクトルの ℓ_2 ノルムが用いられる．一方，スパースな解を導き出す正則化項も最近注目を集めている [1], [10], [14]．例えば，パラメータベクトルの ℓ_1 ノルムを用いると，パラメータベクトルのほとんどが 0 である解（スパースな解）が得られ，特徴選択などに役立つことが知られている．

正則化定数 λ の適切な設定は，良好な汎化性能を得るためには不可欠である．そのための基準として，もっとも理想的なのは汎化誤差そのものであるが，これは実際には計算できないので，汎化誤差の近似法が多く提案されている．例えば，汎化誤差の上限を推定する方法は，VC 理論から導くことができる [16]．また，汎化誤差のアンサンブル平均を推定する方法としては，Network Information Criterion (NIC) [7] や，Subspace Information Criterion (SIC) [12], [13] が提案されている．この他にも，cross validation [16]，C_L [4]，Bayesian evidence framework [3] 等の手法がある．

杉山，小川によって提案された SIC は，汎化誤差のアンサンブル平均の不偏推定量である [12]．SIC の技術的な特徴としては，汎化誤差を推定するために，真のパラメータの不偏推定量を，参照推定量 (reference estimator) として用いる点が挙げられる．また、SIC は、他の統計的モデル選択法と異なり、ラベルのないデータ (unlabeled data) が存在する場合には、この情報を加えて精度を向上させることができる [15]．SIC は、NIC 等の統計的モデル選択法に比べ、サンプル数が少ない場合に優れていることが実験的に示されている [12]．しかし，SIC の欠点は，二次の正則化項を持つ線形回帰問題にしか適用できない点にある．本論文では，SIC を拡張して適用範囲を広げ，ℓ_1 正則化項を持つ線形回帰分析（スパース回帰分析）における正則化定数の設定を行う．この方法を，Generalized Subspace Information Criterion (GSIC) と呼ぶ．GSIC は，あ

[†] 産総研 生命情報科学研究センター, 東京都
AIST Computational Biology Research Center, Aomi 2-41-6, Koto-ku, 135–0064 Japan
[††] 東京工業大学, 東京都
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan
[†††] Fraunhofer FIRST and Potsdam University
Kekuléstr. 7, 12489 Berlin, and Am Neuen Palais 10, 14469 Potsdam, Germany

る仮定のもとで，汎化誤差のアンサンブル平均の漸近不偏推定量であることが示される．

サンプル数がある程度大きい場合には，GSIC は，NIC や cross-validation と同等の性能を示すが，サンプル数が小さい場合には，汎化誤差の推定の分散が大きすぎて，正則化定数の選択が不安定になる．この問題を解決するため，参照推定量として，不偏推定量でなく，二次の正則化項を用いたバイアスを持つ推定量を用いることを考える．これにより，GSIC の漸近不偏性は失われるが，分散を大幅に減少させることができる．この改良版を biased GSIC (GSICb) と呼ぶ．このアプローチの問題点は，参照推定量に関して，正則化定数の設定問題が新たに生じることにある．しかし，二次の正則化項に関しては，Leave-one-out Error [16] が解析的に計算できるので，非常に高速に正則化定数を設定できる．本論文では，小サンプルの場合にも，GSICb が cross validation と同程度の高性能を実現することを示す．また，GSICb の計算量は，参照推定量の正則化定数の選択を含めても，なお cross validation よりも大幅に小さいことも明らかにする．

2. 準　　備

線形回帰の目的は，ノイズを含むデータに，パラメータに関して線形なパラメトリックモデルを当てはめることである．真の関数を $f(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$ とおき，これはパラメトリックモデル $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^{p} \theta_i \phi_i(\boldsymbol{x})$ に含まれているとする．ここで，$\phi_i : \mathbb{R}^d \to \mathbb{R}$ は，固定された基底関数であり，$\boldsymbol{\theta} \in \mathbb{R}^p$ はパラメータベクトルである．真のパラメータを θ_i^* とすると，真の関数は，$f(\boldsymbol{x}) = \sum_{i=1}^{p} \theta_i^* \phi_i(\boldsymbol{x})$ のように表される．訓練サンプルは，入力点 $\boldsymbol{x}_i \in \mathbb{R}^d$ と，それに対応する出力 $y_i \in \mathbb{R}$ から成る．ここでは，出力は加法性ノイズ ϵ_i によって劣化していると仮定する（$y_i = f(\boldsymbol{x}_i) + \epsilon_i$）．ここで，確率変数 ϵ_i は互いに独立で，同じ対称分布に従っていると仮定する．また，この対称分布の平均は 0 で分散は σ^2 である．ここで、σ^2 は既知であるか、または、他のパラメータの推定に影響しないほど高い精度で推定されるとする．パラメータベクトル $\boldsymbol{\theta}$ は，訓練サンプルより，次の損失関数を最小化する解として得られる．

$$L_r = \frac{1}{n} \sum_{i=1}^{n} (f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)^2 + \lambda R(\boldsymbol{\theta}) \qquad (2)$$

ここで，$R(\boldsymbol{\theta})$ は二回微分可能な正則化項であり，λ は正則化定数である．$\hat{\boldsymbol{\theta}}$ を次の最適化問題の解として定義しよう：$\hat{\boldsymbol{\theta}} = \mathrm{argmin}_{\boldsymbol{\theta}} L_r(\boldsymbol{\theta})$．パラメータ $\hat{\boldsymbol{\theta}}$ の汎化誤差は，次のように表される．

$$E_{\boldsymbol{x}}[(f(\boldsymbol{x}) - f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}))^2] = \int (f(\boldsymbol{x}) - f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}))^2 q(\boldsymbol{x}) d\boldsymbol{x} \qquad (3)$$

ここで，$q(\boldsymbol{x})$ は入力分布である．

最適化問題 (2) の解が一意に決まるとすると，$\hat{\boldsymbol{\theta}}$ は，訓練サンプル $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ の陰関数となる：$\hat{\boldsymbol{\theta}}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n, y_1, \cdots, y_n)$．モデル選択においては，正則化定数 λ は，汎化誤差が最小になるように決めるべきである．しかし，$\hat{\boldsymbol{\theta}}$ は，確率変数 y_i に依存しているため，汎化誤差 (3) もまた確率変数である．二つの確率変数の大小を比較するため，ここでは平均のみに注目する．汎化誤差の訓練サンプルに関する平均は，アンサンブル平均と呼ばれ，次のように表される．

$$J_G = E_{\boldsymbol{\epsilon}} E_{\boldsymbol{x}}[(f(\boldsymbol{x}) - f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}))^2] \qquad (4)$$

ここで，$E_{\boldsymbol{\epsilon}} := E_{\epsilon_1} \cdots E_{\epsilon_n}$ である．

汎化誤差のアンサンブル平均を幾何的に理解するため，パラメータ空間の内積を $\langle \boldsymbol{\theta}, \boldsymbol{\theta}' \rangle_P = \boldsymbol{\theta}^T P \boldsymbol{\theta}'$ と定義しよう．ここで，P は，第 (i,j) 要素が $P_{ij} = E_{\boldsymbol{x}}[\phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})]$ である行列である．そうすると，汎化誤差のアンサンブル平均は，パラメータ空間のノルムを用いて

$$J_G = E_{\boldsymbol{\epsilon}} \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_P^2. \qquad (5)$$

と書ける．ここで、$\|\boldsymbol{\theta}\|_P^2 = \langle \boldsymbol{\theta}, \boldsymbol{\theta} \rangle_P$ である。行列 P は入力分布 $q(\boldsymbol{x})$ が既知であれば正確に計算できる．また，$q(\boldsymbol{x})$ が未知の場合でも，訓練サンプルの共分散から推定できる．さらに，y のない unlabeled サンプルが利用できる場合には，これを加えることによって，さらに正確に P が推定できる．

3. 汎化誤差の推定

3.1 基本アイデア

図 1 に基本的なアイデアを示す．前述の通り $\hat{\boldsymbol{\theta}}$ は確率変数なので，パラメータ空間で分布を形づくる（図中の左の楕円）．なお，$\boldsymbol{\theta}^m$ は，$\hat{\boldsymbol{\theta}}$ の平均である：$\boldsymbol{\theta}^m = E_{\boldsymbol{\epsilon}}[\hat{\boldsymbol{\theta}}]$．汎化誤差 J_G は，真のパラメータ $\boldsymbol{\theta}^*$ から $\hat{\boldsymbol{\theta}}$ までの平均距離となるが，我々は $\boldsymbol{\theta}^*$ を知らないので，これを求めることができない．SIC では，ここで参照推定量 $\hat{\boldsymbol{\theta}^u}$ を導入し，これが真のパラメータの不偏推定量であると仮定する：$E_{\boldsymbol{\epsilon}}[\hat{\boldsymbol{\theta}^u}] = \boldsymbol{\theta}^*$．線形回

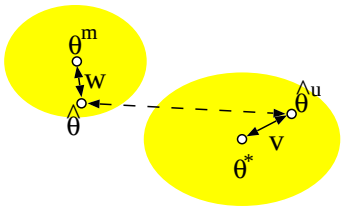


図 1　汎化誤差推定の基本アイデア
Fig. 1　Basic idea for evaluating the generalization error.

帰の場合，最小二乗推定量が真のパラメータの不偏推定量になるので，これを用いれば良い [12]．そうすると，$\hat{\boldsymbol{\theta}}$ と $\hat{\boldsymbol{\theta}^u}$ の距離は，J_G の荒い推定量となるが，これは J_G の不偏推定量ではない．そこで，SIC では修正項を付け加えることで，J_G の不偏推定量を得る．

3.2　一般化 SIC

本節では，J_G の不偏推定量を導きだす．J_G は次のように bias と variance に分解できる．

$$E_{\boldsymbol{\epsilon}}\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_P^2 = \|\boldsymbol{\theta}^m-\boldsymbol{\theta}^*\|_P^2 + E_{\boldsymbol{\epsilon}}\langle \boldsymbol{w},\boldsymbol{w}\rangle_P, \quad (6)$$

ここで，$\boldsymbol{w} := \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^m$ である．bias 項は，$\|\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}^u}\|_P^2$ を用いて次のように変形できる．

$$\|\boldsymbol{\theta}^m-\boldsymbol{\theta}^*\|_P^2 = \|\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}^u}\|_P^2 - \|\boldsymbol{w}-\boldsymbol{v}\|_P^2 - 2\langle \boldsymbol{w}-\boldsymbol{v}, \boldsymbol{\theta}^m-\boldsymbol{\theta}^*\rangle_P, \quad (7)$$

ここで，$\boldsymbol{v} := \hat{\boldsymbol{\theta}^u}-\boldsymbol{\theta}^*$ である．式 7 の第二項と第三項は直接計算できないので，平均で置き換えると，第二項は，

$$-E_{\boldsymbol{\epsilon}}\|\boldsymbol{w}-\boldsymbol{v}\|_P^2 = -E_{\boldsymbol{\epsilon}}\langle \boldsymbol{w},\boldsymbol{w}\rangle_P + 2E_{\boldsymbol{\epsilon}}\langle \boldsymbol{w},\boldsymbol{v}\rangle_P - E_{\boldsymbol{\epsilon}}\langle \boldsymbol{v},\boldsymbol{v}\rangle_P,$$

となり第三項は 0 となる．このような近似によって，次の Generalized Subspace Information Criterion（一般化 SIC）を得る：

$$\mathrm{GSIC} = \|\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}^u}\|_P^2 + 2E_{\boldsymbol{\epsilon}}\langle \boldsymbol{w},\boldsymbol{v}\rangle_P - E_{\boldsymbol{\epsilon}}\langle \boldsymbol{v},\boldsymbol{v}\rangle_P. \quad (8)$$

これは J_G の不偏推定量である．

3.3　二次の正則化項への適用

ここでは，二次の正則化項 $R(\theta) = \boldsymbol{\theta}^T R\boldsymbol{\theta}$ を持つ線形回帰問題に GSIC を適用する．K を (i,j) 要素が $\phi_j(x_i)$ である $n \times p$ 行列であるとする．また，$\boldsymbol{y} = (y_1,\ldots,y_n)^T$ とおく．$(\frac{1}{n}K^TK + \lambda R)$ が正則のとき，$\hat{\boldsymbol{\theta}}$ は次のように与えられる：

$$\hat{\boldsymbol{\theta}} = \frac{1}{n}(\frac{1}{n}K^TK+\lambda R)^{-1}K^Ty. \quad (9)$$

また，K^TK が正則のとき，不偏推定量 $\hat{\boldsymbol{\theta}^u}$ は，次のように書ける：

$$\hat{\boldsymbol{\theta}^u} = (K^TK)^{-1}K^Ty. \quad (10)$$

これら二つの式より，二次正則化項のための SIC を導き出すことができる [12]：

$$\mathrm{SIC} = (\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}^u})^TP(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}^u}) + 2\sigma^2\mathrm{tr}(PW) - \sigma^2\mathrm{tr}(PV), \quad (11)$$

ここで，

$$W = \frac{1}{n}(\frac{1}{n}K^TK+\lambda R)^{-1}, \quad V = (K^TK)^{-1} \quad (12)$$

である．ノイズの分散 σ^2 が未知の場合には，次の推定量を代わりに用いることができる：$n > p$ のとき，$\hat{\sigma}^2 = \{\boldsymbol{y}^T\boldsymbol{y} - (K\hat{\boldsymbol{\theta}^u})^T\boldsymbol{y}\}/(n-p)$ は σ^2 の不偏推定量である [2]．なお，σ^2 を $\hat{\sigma}^2$ を置き換えても，SIC の不偏性には変化はない．しかし、$\hat{\sigma}$ の分散が大きい場合には、SIC の分散が大きくなってしまうので、注意が必要である。

3.4　一般の正則化項への適用

正則化項が二次でないとき，$\hat{\boldsymbol{\theta}}$ は，式 9 のように解析的には得られない．そのため，式 8 の第二項 $E_{\boldsymbol{\epsilon}}\langle \boldsymbol{w},\boldsymbol{v}\rangle_P$ の計算が困難になる．そこで，Hessian 行列 $H = [\frac{\partial^2 L_P}{\partial\theta_i\partial\theta_j}]$ が正則であるという仮定のもとで第二項を近似することを考える：

$$E_{\boldsymbol{\epsilon}}\langle \boldsymbol{w},\boldsymbol{v}\rangle_P \approx \sigma^2\mathrm{tr}(PW^0), \quad (13)$$

ここで，

$$W^0 = \frac{1}{n}\left(\frac{1}{n}K^TK + \frac{1}{2}\lambda\nabla\nabla R(\hat{\boldsymbol{\theta}})\right)^{-1} \quad (14)$$

であり，$\nabla\nabla R(\hat{\boldsymbol{\theta}})$ は，第 (i,j) 要素が $\frac{\partial R(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ であるような行列である．この近似の導出法に関しては，付録を参照して頂きたい．一般の正則化項に対する GSIC は，次のようになる：

$$\mathrm{GSIC} = (\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}^u})^TP(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}^u}) + 2\sigma^2\mathrm{tr}(PW^0) - \sigma^2\mathrm{tr}(PV). \quad (15)$$

正則化項が二次のときには，GSIC は従来の SIC（式 11）と一致する．また，σ^2 が未知のときには，SIC と同様に $\hat{\sigma}^2$ と置き換えればよい．GSIC と J_G には，次の定理で示されるような関係がある．

[定理 1] $\hat{\boldsymbol{\theta}}$ が，$\boldsymbol{y}$ の $b(<\infty)$ 次の多項式で表現でき，ϵ_i の b 次までのモーメントが有限のとき，式 15 の GSIC は，J_G の不偏推定量になる：$E_{\boldsymbol{\epsilon}}[\mathrm{GSIC}] = J_G + O(n^{-2})$.

この定理の証明に関しては [15] を参照して頂きたい.

4. Biased GSIC

訓練サンプル数が非常に小さい状況では，不偏推定量 $\hat{\boldsymbol{\theta}^u}$ の分散が大きくなり，これによって GSIC の分散も非常に大きくなる．このような場合には，正則化定数の選択結果は非常に不安定になる．GSIC の分散を減らすためには，$\hat{\boldsymbol{\theta}^u}$ を，二次の正則化を加えた推定量 $\hat{\boldsymbol{\theta}^\alpha}$ で置き換えるのが効果的である：$\hat{\boldsymbol{\theta}^\alpha} = (K^TK + \alpha I)^{-1}K^T\boldsymbol{y}$, ここで，$I$ は $p \times p$ の単位行列である．参照推定量に正則化を加えたことで GSIC は，J_G の漸近不偏推定量ではなくなるが，分散を格段に小さくできる．この改良版を，biased GSIC (GSICb) と呼ぶ．ここで，新たな正則化定数 α が現れ，これの適切な設定が問題となる．しかし，α を求めるのは，スパース回帰において λ を求めるのに比べて遥かに簡単である．その理由は，二次の正則化項に関しては，Leave-one-out Error が解析的に求まるなど，正則化定数を高速に決定する方法が充実しているからである [9]．例えば，Leave-one-out error は次の式で求めることができる：

$$\mathrm{LOOerror} = \frac{\boldsymbol{y}^TU(\mathrm{diag}(U))^{-2}U\boldsymbol{y}}{n}, \qquad (16)$$

ここで，$U = I_n - K(K^TK + \alpha I)^{-1}K^T$ であり，I_n は，$n \times n$ の単位行列である．なお，ノイズの分散 σ^2 の推定量も，$\hat{\boldsymbol{\theta}^\alpha}$ から次のように求められる [17]：$\hat{\sigma}^2 = \boldsymbol{y}^TZ^2\boldsymbol{y}/\mathrm{tr}(Z)$．ここで，$Z = I - K(K^TK + \alpha I)^{-1}K^T$ である．

5. スパース回帰への適用

ℓ_1 ノルム正則化項を用いれば，ほとんどのパラメータが 0 であるようなスパースな解が得られることが知られている [5], [18]．スパース解は，自動的に特徴を選びだすという点で有用であり，また，計算時間の短縮にも役立つ．スパース回帰分析における損失関数は次のように書ける：

$$L_r = \frac{1}{n}\sum_{i=1}^{n}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)^2 + \lambda\sum_{i=1}^{p}|\theta_i|. \qquad (17)$$

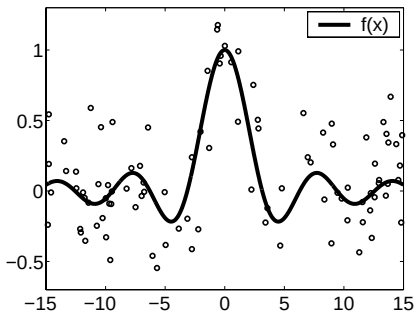


図 2 一次元の人工データにおける真の関数と，100 個の訓練サンプル．$\sigma = 0.3$
Fig. 2 Learning target function and 100 training examples with $\sigma = 0.3$.

$\boldsymbol{\theta}$ の最適化は凸二次計画問題を解くことで行える [6].

スパース回帰では，$\nabla\nabla R$ が定義できないので，GSIC を直接適用することはできない．GSIC を適用するために，正則化項 $R(\boldsymbol{\theta}) = \sum_{i=1}^{p}|\theta_i|$ を次の連続関数で近似する：

$$R'(\boldsymbol{\theta}) = \sum_{i=1}^{p}\theta_i\tanh(\gamma\theta_i), \qquad (18)$$

ここで，$\gamma = 10$ と実験的に設定した．この近似によって，$\nabla\nabla R$ は，

$$\nabla\nabla R_{ii} = 2(\gamma\mathrm{sech}^2(\gamma\hat{\theta}_i) - \gamma^2\hat{\theta}_i\mathrm{sech}^2(\gamma\hat{\theta}_i)\tanh(\gamma\hat{\theta}_i))$$

を要素とする対角行列となる．

6. 実　　験

6.1 人工データによる実験

本節では，一次元の人工データを用いて実験を行う．回帰関数を $f_{\boldsymbol{\theta}}(x) = \sum_{i=1}^{50}\theta_i\exp\left(-\frac{\|x-s_i\|^2}{\eta^2}\right)$, とする．ここで，$\eta = 1$ であり，50 個のテンプレート点 s_i は，$[-15, 15]$ の区間に等間隔に並んでいる．真のパラメータ $\boldsymbol{\theta}^*$ は，関数 $g(x) = |x|^{-1}\sin|x|$ からのサンプル $\{(s_i, g(s_i))\}_{i=1}^{50}$ に回帰関数を最小二乗法であてはめて得た．訓練サンプルを生成するため，n 個の入力点 $\{x_i\}_{i=1}^{n}$ を，区間 $[-15, 15]$ の一様分布からサンプリングした．さらに，出力値は，$y_i = f(x_i) + \epsilon_i$ のように得た．ここで，ϵ_i は，平均 0，分散 σ^2 の正規分布に独立に従う．真の関数と，訓練サンプルの例を図 2 に示した．

スパース回帰分析の正則化定数は，$\lambda = 1.0 \times 10^{-4}, 1.0 \times 10^{-3.5}, \ldots, 1.0 \times 10^{-1}$ から選択される．

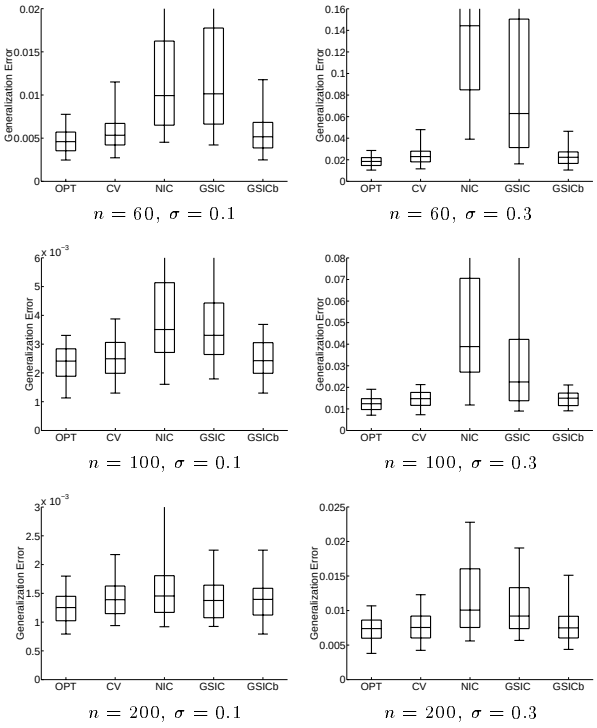


図 3　各基準によって選択された λ における汎化誤差の Box plot. Box plot に含まれる線は，それぞれ 95,75,50,25,5 percentile を示している．ここで，'OPT' は，最適な λ における汎化誤差を示す.

Fig. 3　Generalization errors at selected λ for the respective model selection criterion shown with standard box plot (100 trials). The box plot notation specifies marks at 95, 75, 50, 25, and 5 percentiles of values. 'OPT' denotes the generalization error with the optimal λ.

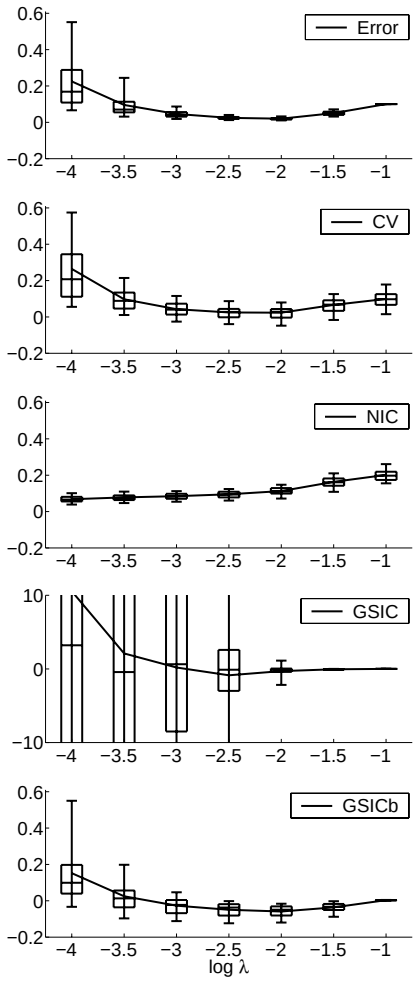


図 4　$n = 60$,$\sigma = 0.3$ における各基準による汎化誤差の推定値．一番上のグラフは真の汎化誤差を表す．異なる訓練セットで 100 回行った結果を Box plot で表示．横軸は $\log\lambda$ を表し，また，図中の実線は平均値を示す.

Fig. 4　Values of each criterion by 100 trials shown with standard box plot ($n = 60$, $\sigma = 0.3$). The horizontal axis denotes $\log\lambda$. The solid line denotes the mean values.

選択に用いた基準は，10-fold cross validation (CV), NIC, GSIC, GSICb である．NIC の細かい設定については、付録 2. を参照して頂きたい．GSIC と GSICb においては，100 個の unlabeled サンプルを区間 $[-15, 15]$ の一様分布からサンプリングし，入力分布 $q(x)$ を，これらを用いて $q(x) = \frac{1}{100}\sum_{i=1}^{100}\delta(x - x_i')$ とおいた．ここで，$\delta(x)$ は，$x = 0$ のとき 1 となり，その他の場合には 0 となるデルタ関数である．これにより、汎化誤差は、この 100 点においてのみ計測されることになるが、このような低次元データにおいては、100 点における計測でも十分正確なモデル選択が行えると考えた。

パラメータ $\boldsymbol{\theta}$ の汎化誤差は，次のように表される：$\text{Error} = \int_{-15}^{15}(f_{\boldsymbol{\theta}}(x) - f(x))^2\,dx$．各手法の性能は，選択された λ における解の汎化誤差で評価する（図 3）．実験は，訓練セットを変化させて 100 回行った．$n = 200$ のときには，すべての基準が同様の性能を示している．しかし，n が 60 に減少すると，CV は高い性能を保つが，NIC と GSIC は大きく悪化してしまう．

NIC と GSIC の推定誤差の原因を知るため，汎化誤差の推定値を図 4,5 に示す．この図において，GSIC

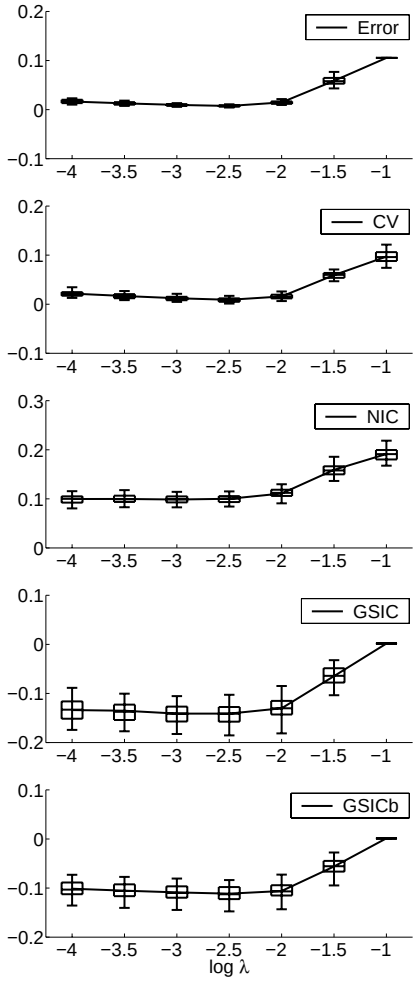


図 5　$n = 200, \sigma = 0.3$ における各基準による汎化誤差の推定値．各軸の意味等は図 4 参照

Fig. 5　Values of each criterion by 100 trials shown with standard box plot ($n = 200$, $\sigma = 0.3$).

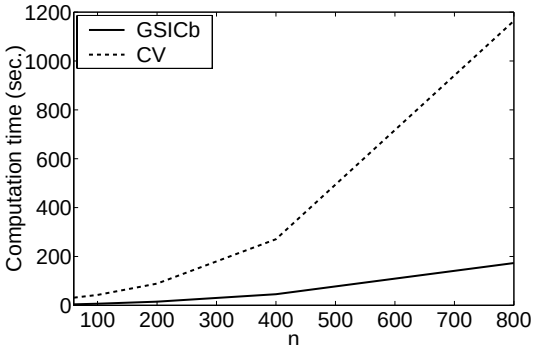


図 6　Cross validation(CV) と GSICb の計算時間の比較．横軸は訓練サンプル数，縦軸は計算時間（秒）を示す

Fig. 6　Computation time. The horizontal axis denotes the number of training examples and the vertical axis denotes the computation time in seconds.

の値にはバイアスがあるが，これはモデル選択に無関係な項 $\|\hat{\boldsymbol{\theta}}^u\|^2$ と，$\sigma^2 \mathrm{tr}(PV)$ を除いたからである．これによって，モデル選択に本質的な分散を観測できる．$n = 200$ の場合 (図 5) には，CV,NIC,GSIC の曲線は，ともに真の汎化誤差に近い．しかし，$n = 60$ の場合 (図 4) には，CV が依然として近い曲線を示すのに対し，NIC と GSIC は正確でなくなる．NIC と GSIC の曲線を見てみると，二つの手法の推定誤差は異なる特徴を示している．NIC のカーブは左に傾いており，最適なものよりも小さな正則化定数を選ぶ傾向があることを示している．NIC の平均は，正しい曲線を大きくはずれているので，小サンプル効果によって不偏性が失われてしまっているといえる．一方，GSIC の曲線では，特に λ が小さい部分で，巨大な分散が全体を圧していることがわかる．まとめて言えば，NIC の誤差ではバイアスが主要な役割を果たしているのに対し，GSIC では分散の方がより大きな問題であることが分かる．

GSIC の分散を減少させるため，GSICb を用いることができる．ここでは，参照推定量の正則化定数 α を決定するのに，Leave-one-out Error（式 16）を用いて，$\alpha = 1.0 \times 10^{-4}, 1.0 \times 10^{-3.5}, \ldots, 1.0 \times 10^{1}$ から選択した．図 3 によると，GSICb は，$n = 200$ のときは他の手法と同様の性能を示すが，$n = 60$ になってもあまり悪化せず CV と同様の高い性能を維持することができる．また，図 4 から，GSICb の曲線は，サンプル数が少ないときでも，真の汎化誤差に近いことがわかる．特に，GSICb の分散は，GSIC に比べ大幅に減少しているのに対し，バイアスの増加の方は顕著でない．

GSICb と CV の計算時間の比較を図 6 に示す．ここで，GSICb のグラフは，α の選択に要した時間も含めた全計算時間を表す．また，CV のグラフは，実際に凸二次計画問題を繰り返し解いて 10-fold cross validation を行うのに要した計算時間を示す．GSICb の計算時間の方が遥かに短く，またその差は n が大きくなるにつれて拡大することがわかる．

6.2　実データでの実験

GSICb の実データでの有効性を確かめるため，

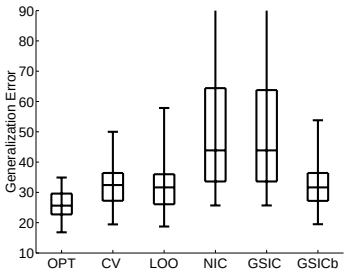


図 7　Boston housing data における各基準によって選択された λ における汎化誤差の Box plot．'LOO' は，leave-one-out cross validation の結果を示す

Fig. 7　Generalization errors at selected λ by 100 trials in Boston Housing data. The results are shown with the standard box plot. 'LOO' denotes the result of leave-one-out cross validation

DELVE データベース [8] に含まれている Boston housing data を用いて実験を行う．このデータは 14 次元空間に 506 個の点を含むが，ここでは，第 14 次元の MEDV を出力とし，他の次元を入力とした．各入力次元の値は，最大値で割ることによって正規化した．ランダムに選んだ 50 点を訓練サンプルとし，100 個を unlabeled サンプルとした．残りの 356 個はテストサンプルとし，汎化誤差の測定に用いた．本実験での回帰関数は，次のように表される．

$$f_{\boldsymbol{\theta}}(x) = \sum_{i=1}^{50} \theta_i k(\boldsymbol{x}, \boldsymbol{x}_i) \qquad (19)$$

ここで，k は，3 次の linear spline ANOVA カーネル [11], [16] であり，50 個の訓練サンプルすべてがテンプレート点として用いられている．正則化定数は，区間 $[10^{-3}, 10^{3}]$ から log スケールで等間隔に 10 点を取り，その中から選ぶ．なお，GSICb の α も同じ 10 点の中から選んだ．訓練セットを変化させて 100 回正則化定数選択実験を行った結果を図 7 に示す．ここでは，10-fold cross validation に加えて，leave-one-out cross validation の結果も示した．この実験のように，訓練サンプルの数がパラメータ数と等しい場合にも，GSICb は，cross validation と同様の性能を示すことがわかる．また、各基準の推定値を図 8 に示す。ここでは、GSIC が NIC と同様に大きなバイアスを持っているが、これは一般のデータにおいては、最初の仮定のようにモデルに真の関数が含まれないためだと考えられる。

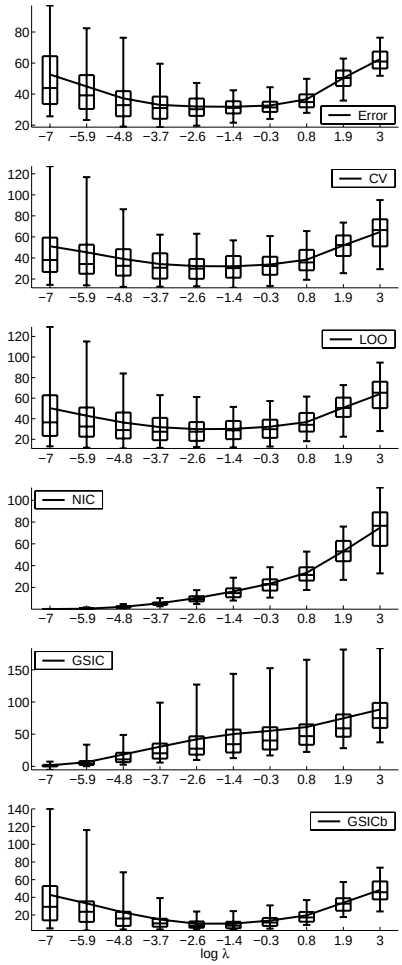


図 8　Boston housing data における各基準による汎化誤差の推定値．各軸の意味等は図 7 参照

Fig. 8　Values of each criterion by 100 trials in Boston housing data.

7. おわりに

本論文では，SIC を拡張し，GSIC と GSICb という二つの手法を導き出して，それらをスパース回帰分析に適用した．理論的には，GSIC のある条件のもとでの漸近不偏性を示し，実験においては，GSICb が，小サンプルの場合にも，うまく対応できることを示した．今後の課題としては，GSICb に対しての理論的解析と，教師なし学習への，GSIC(b) の適用が挙げられる．

謝辞　有益な議論をして頂いた Shun-ichi Amari, Hidemitsu Ogawa, Takashi Onoda, Gunnar Rätsch, Sebastian Mika の各氏に感謝致します．K.R.M grate-

fully acknowledges partial financial support from DFG under contracts JA 379/91, MU 987/11 and the EU in the Neurocolt 2 and the BLISS project (IST-1999-14190).

文　　献

[1] K.P. Bennett and O.L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–34, 1992.

[2] V.V. Fedorov. *Theory of Optimal Experiments*. Academic Press, New York, 1972.

[3] D.J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[4] C.L. Mallows. Some comments on C_P. *Technometrics*, 15(4):661–675, 1973.

[5] O.L. Mangasarian. Mathematical programming in data mining. *Data Mining and Knowledge Discovery*, 42(1):183–201, 1997.

[6] S. Mika, G. Rätsch, and K.-R. Müller. A mathematical programming approach to the kernel fisher algorithm. to appear in Neural Information Processing Systems 13, 2001.

[7] N. Murata, S. Yoshizawa, and S. Amari. Network information criterion — determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, 5:865–872, 1994.

[8] University of Toronto. http://www.cs.utoronto.ca/~delve/data/datasets.html. DELVE-Benchmark repository – a collection of artificial and real-world data sets.

[9] M.J.L. Orr. Introduction to radial basis function networks. Technical report, Centre for Cognitive Science, University of Edinburgh, 1996.

[10] V. Roth. Sparse kernel regressors. *Proc. ICANN'01*, pages 339–346, 2001.

[11] M. Stitson, A. Gammerman, V.N. Vapnik, V. Vovk, C. Watkins, and J. Weston. Support vector regression with anova decomposition kernels. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 285–291. MIT Press, Cambridge, MA, 1999.

[12] M. Sugiyama and H. Ogawa. Subspace information criterion for model selection. *Neural Computation*, 13(8), 2001.

[13] M. Sugiyama and H. Ogawa. Theoretical and experimental evaluation of subspace information criterion. *Machine Learning*, 2001. to appear.

[14] R.J. Tibshirani. Regression shrinkage and selection via the LASSO. *J. Royal Statistical Society B*, 58(1):267–288, 1996.

[15] K. Tsuda, M. Sugiyama, and K.-R. Müller. Subspace information criterion for non-quadratic regularizers. Technical Report 120, GMD, 2000. http://wsv.gmd.de/aiw/report.htm.

[16] V.N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[17] G. Wahba. *Spline Model for Observational Data*, volume 59 of *Series in Applied Mathematics*. SIAM: Society for Industrial and Applied Mathematics, Philadelphia and Pennsylvania, 1990.

[18] P.M. Williams. Bayesian regularisation and pruning using a Laplace prior. *Neural Computation*, 7(1):117–143, 1995.

付　　録

1. GSIC の導出

この章では，式 13 の導出を行う．学習の最適化問題の解が一意だと仮定すると，$\hat{\boldsymbol{\theta}}$ は $\boldsymbol{y}$ の関数となる．$\boldsymbol{z} := (f(x_1), \ldots, f(x_n))^T$，$\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_n)^T$ とおくと，$\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ の微分は，次のように書ける：

$$\nabla\hat{\theta}_i(\boldsymbol{y}) := \left(\frac{\partial\hat{\theta}_i}{\partial y_1}(\boldsymbol{y}), \ldots, \frac{\partial\hat{\theta}_i}{\partial y_n}(\boldsymbol{y})\right)^T, \quad \text{(A·1)}$$

ここで，$\nabla\hat{\boldsymbol{\theta}}(\boldsymbol{y}) := (\nabla\hat{\theta}_1(\boldsymbol{y}), \cdots, \nabla\hat{\theta}_p(\boldsymbol{y}))^T$ である．さらに，$\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ をテイラー展開すると，

$$\hat{\theta}_i(\boldsymbol{y}) = \hat{\theta}_i(\boldsymbol{z}+\boldsymbol{\epsilon}) = \hat{\theta}_i(\boldsymbol{z}) + \nabla\hat{\theta}_i(\boldsymbol{z})^T\boldsymbol{\epsilon} + S_i, \quad \text{(A·2)}$$

ここで，S_i は剰余項である．これより，$\boldsymbol{w}$ の第 i 要素 w_i は，

$$w_i = \hat{\theta}_i(\boldsymbol{z}+\boldsymbol{\epsilon}) - E_{\boldsymbol{\epsilon}}[\hat{\theta}_i(\boldsymbol{z}+\boldsymbol{\epsilon})] = \nabla\hat{\theta}_i(\boldsymbol{z})^T\boldsymbol{\epsilon} + S_i - E_{\boldsymbol{\epsilon}}[S_i].$$

と書ける．不偏推定量 $\hat{\boldsymbol{\theta}^u}(\boldsymbol{y})$ を式 10 のように定義すると，

$$\nabla\hat{\boldsymbol{\theta}^u} = (K^TK)^{-1}K^T \quad \text{(A·3)}$$

これより，$\boldsymbol{v}$ の第 i 要素 v_i は，

$$v_i = \nabla\hat{\theta}_i^{uT}\boldsymbol{\epsilon} \quad \text{(A·4)}$$

と表される．これらを用いると，$E_{\boldsymbol{\epsilon}}\langle\boldsymbol{w},\boldsymbol{v}\rangle_P$ は，

$$E_{\boldsymbol{\epsilon}}\langle\boldsymbol{w},\boldsymbol{v}\rangle_P = \sum_{i=1}^{p}\sum_{j=1}^{p} P_{ij}E_{\boldsymbol{\epsilon}}[w_iv_j] \quad \text{(A·5)}$$

となる．ここで，

$$E_{\boldsymbol{\epsilon}}[w_iv_j] = \sigma^2\nabla\hat{\theta}_i(\boldsymbol{z})^T\nabla\hat{\theta}_j^u + E_{\boldsymbol{\epsilon}}[S_i(\nabla\hat{\theta}_j^{uT}\boldsymbol{\epsilon})]. \quad \text{(A·6)}$$

(A·6) の第一項は、漸近展開によって次のように書き換えられる：

$$\sigma^2 \nabla\hat{\theta}_i(\boldsymbol{z})^T \nabla\hat{\theta}_j^u = \sigma^2 \nabla\hat{\theta}_i(\boldsymbol{y})^T \nabla\hat{\theta}_j^u + O(n^{-2}) \tag{A·7}$$

また、(A·6) の第二項は、次のようになる：

$$E_{\boldsymbol{\epsilon}}[S_i(\nabla\hat{\theta}_j^{u\,T}\boldsymbol{\epsilon})] = O(n^{-2}) \tag{A·8}$$

漸近展開の詳細に関しては, [15] を参照して頂きたい. サンプル数の大きいときには、$O(n^{-2})$ を無視できるので、$E_{\boldsymbol{\epsilon}}\langle \boldsymbol{w}, \boldsymbol{v}\rangle_P$ は次のように表される：

$$E_{\boldsymbol{\epsilon}}\langle \boldsymbol{w}, \boldsymbol{v}\rangle_P \approx \sigma^2 \mathrm{tr}(PW^0), \quad W^0 = \nabla\hat{\boldsymbol{\theta}}(\boldsymbol{y}) \nabla\hat{\boldsymbol{\theta}}^{u\,T}. \tag{A·9}$$

ここに含まれる微分 $\nabla\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ は，鞍点方程式

$$\frac{\partial L_r}{\partial \theta_i} = 0, \ (i = 1, \ldots, p). \tag{A·10}$$

を解いて得る．この方程式の両辺を y_k で微分すると，$H\nabla\hat{\boldsymbol{\theta}}(\boldsymbol{y}) + M = \boldsymbol{0}$．ここで，$H$ は，(i,j) 要素が，$H_{ij} = \frac{\partial^2 L_r}{\partial\theta_i \partial\theta_j}$ である $p \times p$ 行列であり，M は，(i,j) 要素が $M_{ij} = \frac{\partial^2 L_r}{\partial\theta_i \partial y_j}$ である $p \times n$ 行列である．仮定から，H は正則なので，$\nabla\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ は，次のように表せる：

$$\nabla\hat{\boldsymbol{\theta}}(\boldsymbol{y}) = -H^{-1}M. \tag{A·11}$$

式 2 を式 A·11 に代入すると，

$$\nabla\hat{\boldsymbol{\theta}}(\boldsymbol{y}) = \frac{1}{n}\left(\frac{1}{n}K^TK + \frac{1}{2}\lambda\nabla\nabla R(\hat{\boldsymbol{\theta}}(\boldsymbol{y}))\right)^{-1} K^T. \tag{A·12}$$

従って，式 14 は，式 A·3 と式 A·12 を式 A·9 に代入して得ることができる．

2. NIC について

NIC の定義式は、損失関数を $d(\boldsymbol{x}, y, \hat{\boldsymbol{\theta}})$ と置くと、

$$\mathrm{NIC}(\hat{\boldsymbol{\theta}}) = E_t[d(\boldsymbol{x}, y, \hat{\boldsymbol{\theta}})] + \frac{1}{n}\mathrm{tr}(G(\hat{\boldsymbol{\theta}})Q(\hat{\boldsymbol{\theta}})^{-1})$$

である [7]。ここで、

$$G(\hat{\boldsymbol{\theta}}) = V_t[\nabla_{\boldsymbol{\theta}} d(\boldsymbol{x}, y, \hat{\boldsymbol{\theta}})]$$
$$Q(\hat{\boldsymbol{\theta}}) = E_t[\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} d(\boldsymbol{x}, y, \hat{\boldsymbol{\theta}})].$$

なお、E_t, V_t は、それぞれ訓練サンプルに関する期待値、分散であり、$\nabla_{\boldsymbol{\theta}}$ は、$\boldsymbol{\theta}$ の各要素に関する偏微分を表す。損失関数は、近似された正則化項 (18) を用いて、

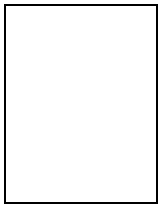
$$d(\boldsymbol{x}, y, \boldsymbol{\theta}) = (f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y)^2 + \lambda\sum_{i=1}^{p}\theta_i \tanh(\gamma\theta_i) \tag{A·13}$$

とおいた。NIC は、損失関数の真の期待値 $E[d(\boldsymbol{x}, y, \hat{\boldsymbol{\theta}})]$ の推定量となるが、これをそのまま用いると、汎化誤差の他に正則化項が含まれてしまうので、ここでは、

$$\mathrm{NIC}(\hat{\boldsymbol{\theta}}) - \lambda\sum_{i=1}^{p}\hat{\theta}_i \tanh(\gamma\hat{\theta}_i) \tag{A·14}$$

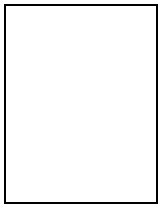
を汎化誤差の推定量として用いた。

（平成 x 年 xx 月 xx 日受付）



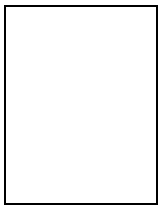
津田　宏治　（正員）

1998 京都大学工学研究科情報工学専攻博士課程了．現在，産業技術総合研究所生命情報科学研究センター研究員．機械学習とカーネル法の研究に従事．



杉山　　将　（正員）

2001 東京工業大学大学院情報理工学研究科計算工学専攻博士課程了．現在，同専攻助手．機械学習の理論研究，及び，その画像処理，パターン認識への応用研究に従事．



Klaus-RobertMüller

1992University of Karlsruhe において理論計算機科学の Ph.D 取得．1992 から 1994 までベルリンの GMD FIRST にて Postdoctoral fellow，IDA(intelligent data analysis) グループの立ち上げを行う．1994 から 1995 まで，東大甘利研に European Community STP Research Fellow として滞在．1995 より GMD FIRST の department head．1999 より GMD と Potsdam 大学の joint associate Professor となる．1999German pattern recognition society (DAGM) より，annual national prize for pattern recognition (Olympus Prize) を受賞．Computational Statistics，IEEE Trans. Biomedical Engineering の編集委員を務める一方，数々の国際会議の organizing commitee にも加わる．研究分野は，統計物理，ニューラルネットのための統計的学習理論，サポートベクターマシン，アンサンブル学習など多岐にわたる．最近の興味は，医学的データ分析のための，時系列分析，Blind source separation，statistical denoising にも広がっている．