日本神経回路学会誌, vol.13, no.3, pp.111–118, 2006.

共変量シフト下での教師付き学習

杉山 将

東京工業大学 計算工学専攻

〒152-8552 東京都目黒区大岡山 2-12-1-W8-74

sugi@cs.titech.ac.jp

http://sugiyama-www.cs.titech.ac.jp/~sugi

概 要

これまで，教師付き学習は訓練用の標本がテスト時に用いる標本と同じ規則に従って生成されるという大前提のもとで研究されてきた．しかし，現実的な場面ではこの前提が成り立たないことも多く，そのような場合は従来の教師付き学習法では良い学習結果が得られない．このような背景のもと，共変量シフトと呼ばれる状況下での学習法が近年盛んに研究されている．共変量シフトとは，与えられた入力に対する出力の生成規則は訓練時とテスト時で変わらないが，入力（共変量）の分布が訓練時とテスト時で異なるという状況である．本稿では，共変量シフト下での教師付き学習の最近の研究成果を概説する．

1 はじめに

教師付き学習は，入力と出力の組から成る訓練標本を用いて，将来与えられるテスト入力に対する出力を予測する問題である．この時，訓練標本とテスト標本が**同じ**規則に従って生成されるというのが大前提である [40, 38, 10, 25]．

しかし，現実的な場面ではこの前提が満たされないことも多い．例えば非定常なシステムでは，標本の生成規則が時々刻々と変化する．このような場合，昔に採取した訓練標本は破棄して，比較的最近に集めた標本のみを使う忘却型の学習アルゴリズムがよく用いられる [20, 18]．標本生成のメカニズムが徐々に変化する場合は，このような方法によりテスト出力を効果的に予測することができる．

一方，標本の生成規則が訓練時とテスト時で大きく変化する場合，忘却型のアルゴリズムではうまく対応できない．もちろん，訓練標本とテスト標本が全く異なる規則により生成された場合は，訓練標本からテスト標本に関する情報が全く抽出できない．従って，意味のある議論をするためには，訓練時とテスト時の標本に何らかの共通点が必要である．

入出力規則（与えられた入力に対する出力の生成規則）は訓練時とテスト時で変わらないが，入力（共変量）の分布が訓練時とテスト時で異なるという状況は**共変量シフト**と呼ばれている [28]．訓練入力がほとんど無い場所で予測を行なう外挿問題は共変量シフトの典型的な例であろう．また，訓練入力の場所をユーザが決定する能動学習 [7, 21] では，共変量シフトが必然である．近年盛んに研究されているブレイン・コンピュータインターフェース [43]，方策オフ型強化学習 [27]，バイオインフォマティックス [4] などでも，共変量シフトが見受けられる．また，経済学における標本選択バイアス [11] も，共変量シフトの一形態とみなすことができるであろう．

本稿では，共変量シフト下での学習法を紹介する．

2 定式化

本節では，共変量シフト下での教師付き学習問題を定式化する．

訓練標本を $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ で表す．但し，訓練入力 $\boldsymbol{x}_i \in \mathcal{D}_{\boldsymbol{x}} \subset \mathbb{R}^d$ は確率分布 $P_0(\boldsymbol{x})$ に独立に従い，訓練出力 $y_i \in \mathcal{D}_y \subset \mathbb{R}$ は条件付き確率分布 $P(y|\boldsymbol{x})$ に独立に従うと仮定する．ここで，$P(y|\boldsymbol{x})$ の条件付き期待値を $f(\boldsymbol{x})$ で表し，条件付き分散は $\boldsymbol{x}$ に依らず一定値 σ^2 であると仮定する．このとき $P(y|\boldsymbol{x})$ は，真の出力 $f(\boldsymbol{x})$ に平均 0，分散 σ^2 の独立な雑音が加わったものと解釈できる．また，訓練入力と訓練出力の集合をそれぞれ $\mathcal{X}_n = \{\boldsymbol{x}_i\}_{i=1}^n$ と $\mathcal{Y}_n = \{y_i\}_{i=1}^n$ で表す．

損失関数 $\ell(\boldsymbol{x}, y, \widehat{y}) : \mathcal{D}_{\boldsymbol{x}} \times \mathcal{D}_y \times \mathcal{D}_y \to [0, \infty)$ は，入力 $\boldsymbol{x}$ での本当の出力 y を $\widehat{y}$ と推定したときの損失を表す．$\mathcal{D}_y$ が実数集合である回帰問題の場合は，二乗損失

$$\ell(\boldsymbol{x}, y, \widehat{y}) = (\widehat{y} - y)^2 \tag{1}$$

がよく用いられる．一方，$\mathcal{D}_y$ が離散集合である分類問題の場合は，誤分類確率に対応する 0/1-損失

$$\ell(\boldsymbol{x}, y, \widehat{y}) = \begin{cases} 0 & \widehat{y} = y \text{ の時} \\ 1 & \text{それ以外の時} \end{cases} \tag{2}$$

がよく用いられる．

入出力関係の学習に，パラメータ $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^b$ を持つモデル（関数族）$\widehat{f}(\boldsymbol{x}; \boldsymbol{\theta})$ を用いる．ここで，モデルが正しいことは仮定しない．即ち，真の関数 $f(\boldsymbol{x})$ が上記のモデルに含まれているとは限らない．教師付き学習の目的は，学習に用いていないテスト入力に対する出力を正しく予測できるように，パラメータ $\boldsymbol{\theta}$ を決定することである．テスト標本を $(\boldsymbol{t}, u)$ で表す．$\boldsymbol{t} \in \mathcal{D}_{\boldsymbol{x}}$ はテスト入力で，$u \in \mathcal{D}_y$ はテスト出力である．テスト出力 u は，訓練出力と同じ条件付き確率分布 $P(u|\boldsymbol{t})$ に従うと仮定する．このテスト標本を用いて，次の**汎化誤差**（期待テスト誤差） G を最小にするようにパラメータ $\boldsymbol{\theta}$ を学習することを目指す．

$$G = \mathop{\mathbb{E}}_{\boldsymbol{t}, u} \ell(\boldsymbol{t}, u, \widehat{f}(\boldsymbol{t}; \boldsymbol{\theta})) \tag{3}$$

但し，$\mathbb{E}$ は期待値を表す．

標準的な教師付き学習の枠組みでは，テスト入力 $\boldsymbol{t}$ は訓練入力 $\{\boldsymbol{x}_i\}_{i=1}^n$ と同じ確率分布 $P_0(\boldsymbol{x})$ に従うと仮定する [40, 38, 10, 25]．しかし，本稿では**共変量シフト**[28] の状況，即ち，テスト入力 $\boldsymbol{t}$ は訓練入力とは異なる確率分布 $P_1(\boldsymbol{x})$ に従う場合を考える．以下では，訓練入力とテスト入力の確率密度関数をそれぞれ $p_0(\boldsymbol{x})$ と $p_1(\boldsymbol{x})$ で表し，それらが既知かつ非零であると仮定する．

3 パラメータ学習法

汎化誤差 G は未知の期待値 $\mathbb{E}_{\boldsymbol{t},u}$ を含むため，直接これを最小にするようにパラメータを決定することはできない．そこで期待値を訓練標本 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ による平均に置き換えた**経験リスク最小化法**（ERM; Empirical Risk Minimization）[38, 25] がよく用いられる．

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[\frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{x}_i, y_i, \widehat{f}(\boldsymbol{x}_i;\boldsymbol{\theta}))\right] \tag{4}$$

$P_0(\boldsymbol{x}) = P_1(\boldsymbol{x})$ のとき，ERM は適当な正則条件の下で一致性を持つ．即ち，訓練標本数 n を増やしていくと，ERM によって得られる推定量は最適なパラメータ $\boldsymbol{\theta}^*$ に収束する．

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} G \tag{5}$$

しかし，$P_0(\boldsymbol{x}) \neq P_1(\boldsymbol{x})$ のとき ERM はもはや一致性を持たない[1]．

共変量シフト下では，次の**重要度重み付き ERM**（IWERM; Importance Weighted ERM）[29, 19, 28] が一致性を持つ．

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[\frac{1}{n}\sum_{i=1}^{n} \frac{p_1(\boldsymbol{x}_i)}{p_0(\boldsymbol{x}_i)}\ell(\boldsymbol{x}_i, y_i, \widehat{f}(\boldsymbol{x}_i;\boldsymbol{\theta}))\right] \tag{6}$$

$p_1(\boldsymbol{x}_i)/p_0(\boldsymbol{x}_i)$ を**重要度**と呼ぶ．IWERM が一致性を持つことは，大数の法則から直ちに分かる：

$$\begin{aligned}
&\lim_{n\to\infty} \left[\frac{1}{n}\sum_{i=1}^{n} \frac{p_1(\boldsymbol{x}_i)}{p_0(\boldsymbol{x}_i)}\ell(\boldsymbol{x}_i, y_i, \widehat{f}(\boldsymbol{x}_i;\boldsymbol{\theta}))\right] \\
&\quad = \iint \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})}\ell(\boldsymbol{x}, y, \widehat{f}(\boldsymbol{x};\boldsymbol{\theta}))p_0(\boldsymbol{x})p(y|\boldsymbol{x})d\boldsymbol{x}dy \\
&\quad = \iint \ell(\boldsymbol{x}, y, \widehat{f}(\boldsymbol{x};\boldsymbol{\theta}))p_1(\boldsymbol{x})p(y|\boldsymbol{x})d\boldsymbol{x}dy \\
&\quad = \mathop{\mathbb{E}}_{\boldsymbol{t},u} \ell(\boldsymbol{t}, u, \widehat{f}(\boldsymbol{t};\boldsymbol{\theta}))
\end{aligned} \tag{7}$$

[1] 但し，モデルが正しい場合は共変量シフト下でも ERM は一致性を持つ．

これは，**重点サンプリング**[8] で用いられているテクニックを教師付き学習問題に応用したものである．

IWERM を用いることによって，共変量シフトが起こっている場合でも一致性を保証することができる．しかし，IWERM は ERM よりも分散が大きいことが多い．従って，訓練標本数が有限である現実的な場面では必ずしも IWERM が最良な推定法とは限らない．有限標本に対しては，IWLS を安定化させた**適応 IWERM**（AIWERM; Adaptive IWERM）[28] の方が良いであろう．

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{p_1(\boldsymbol{x}_i)}{p_0(\boldsymbol{x}_i)} \right)^{\lambda} \ell(\boldsymbol{x}_i, y_i, \widehat{f}(\boldsymbol{x}_i; \boldsymbol{\theta})) \right] \tag{8}$$

但し，$\lambda \in [0, 1]$ である．$\lambda = 0$ の時 AIWERM は通常の ERM と一致し，$\lambda = 1$ の時 AIWERM は IWERM と一致する．従って，AIWERM は ERM と IWERM の中間の性質を持つ．

ここで，例として入力が一次元の簡単な回帰問題を考えよう．訓練入力とテスト入力の確率密度関数を図 1-(A) に，学習したい真の関数 $f(x)$ と学習結果 $\widehat{f}(x)$ を図 1-(B)–(D) にそれぞれ示す．これは，真の関数（sinc 関数）を直線モデルでフィットする問題である [33, 32]．訓練標本は主に図の左側からしか採取できず，テスト標本は右側にある．従って，これは外挿問題である．図中の G はそれぞれの学習結果の汎化誤差を表す．

図 1-(B) は，**最小二乗法**（OLS; Ordinary Least Squares, 二乗損失を用いた ERM）による学習結果である．OLS は訓練出力（左側）を非常に良く近似しているが，テスト出力（右側）の良い推定ではない．図 1-(D) は，**重要度重み付き最小二乗法**（IWLS, 二乗損失を用いた IWERM）による学習結果である．IWLS は右側のテスト標本にうまく適合していることが分かる．しかし，分散が大きいため学習結果は不安定である．この事は，線形モデルに対する OLS が最良線形不偏推定量であること，即ち，不偏な線形推定量の中で分散が最も小さいことから直感的に理解できるであろう．図 1-(C) は，$\lambda = 0.5$ の**適応重要度重み付き最小二乗法**（AIWLS, 二乗損失を用いた AIWERM）による学習結果である．分散を軽減することにより，OLS や IWLS よりも良い学習結果が得られている．

4 モデル選択法

前節では，AIWERM が共変量シフトにうまく対応できる学習法であることを説明した．しかし，AIWERM には重みを調整する超パラメータ λ が含まれており，これを適切に決定しなければ良い学習結果が得られない．また，学習に用いるモデル（関数族）を適切に決定することも重要である．超パラメータやモデルを決定する問題は**モデル選択**と呼ばれる．標準的なモデル選択法では，未知の汎化誤差の推定量を求め，推定した汎化誤差を最小にするように超パラメータやモデルを決定する．従って，モデル選択研究の本質は如何に精度の良い汎化誤差推定量を構成できるかという点にある．

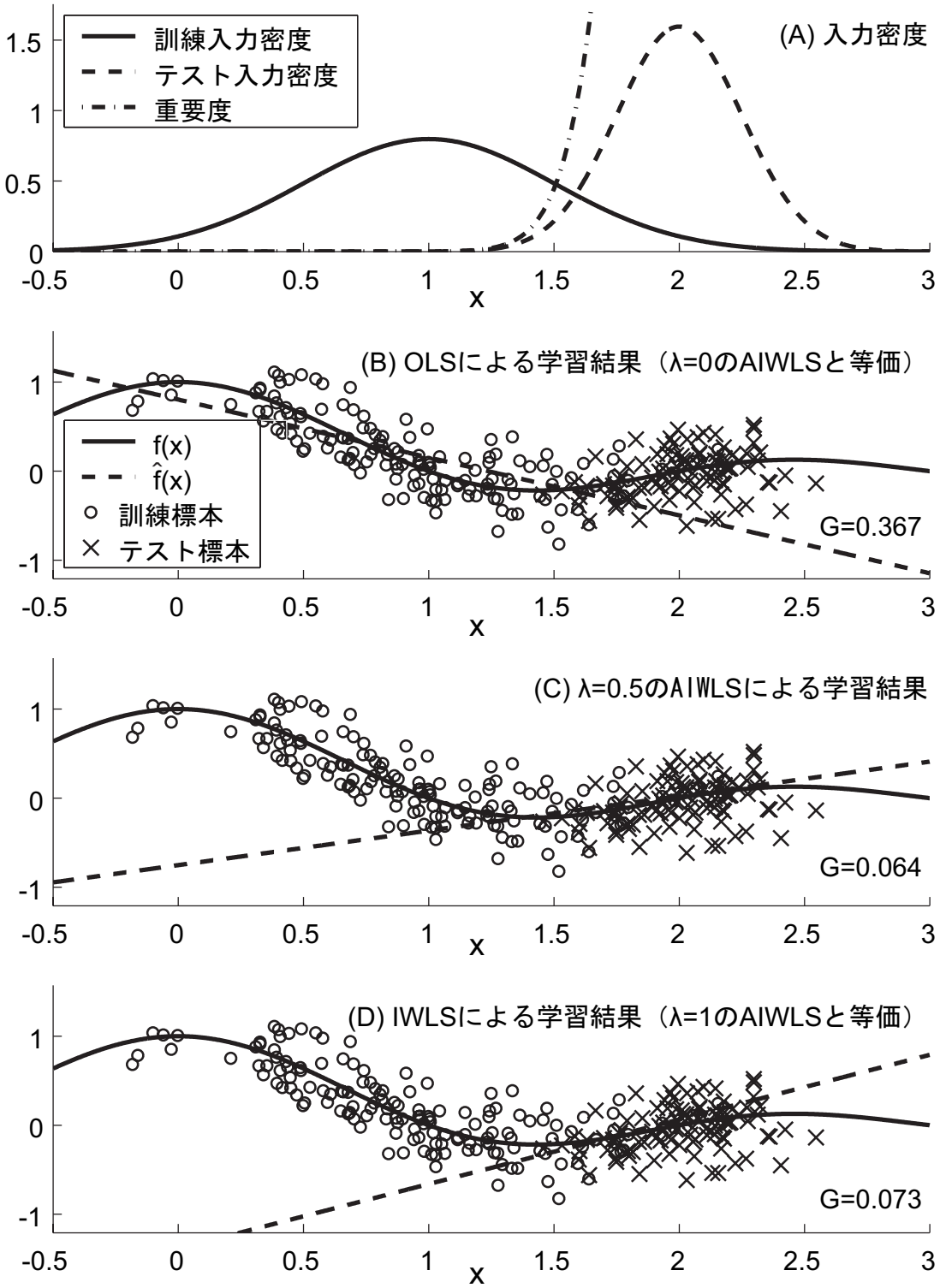

図 1: 共変量シフト下での回帰問題の例．(A) 訓練入力とテスト入力の確率密度関数．(B)–(D) 学習したい真の関数 $f(x)$（実線），訓練標本（‘$\circ$’），学習結果の関数 $\widehat{f}(x)$（破線），テスト標本（‘$\times$’）．

4.1 重要度重み付き交差確認法

これまでに，様々な汎化誤差推定法が提案されてきた．[17, 1, 23, 26, 2, 40, 36]．このうち，最もよく用いられている汎化誤差推定法は**交差確認法**（CV; Cross Validation）[40] であろう．CV では，訓練標本集合 $\mathcal{T} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ を k 個の重なりの無い部分集合 $\{\mathcal{T}_i\}_{i=1}^k$ に分ける．そして，$\{\mathcal{T}_i\}_{i\neq j}$ を用いて学習結果 $\widehat{f}_{\mathcal{T}_j}(\boldsymbol{x})$ を求め，学習に用いなかった $\mathcal{T}_j$ を用いて $\widehat{f}_{\mathcal{T}_j}(\boldsymbol{x})$ の汎化誤差を確認する．これを全ての組み合わせについて平均したものを，$\widehat{f}(\boldsymbol{x})$ の汎化誤差の推定値とする．これを k **重 CV** と呼ぶ．

$$\widehat{G}_{kCV} = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{|\mathcal{T}_j|}\sum_{(\boldsymbol{x},y)\in\mathcal{T}_j}\ell(\boldsymbol{x}, y, \widehat{f}_{\mathcal{T}_j}(\boldsymbol{x})) \tag{9}$$

ここで，$|\mathcal{T}_j|$ は $\mathcal{T}_j$ に含まれる標本数である．$k = n$ のとき，特に**一つ抜き CV**（LOOCV; Leave-One-Out CV）と呼ぶ．

$$\widehat{G}_{LOOCV} = \frac{1}{n}\sum_{j=1}^{n}\ell(\boldsymbol{x}_j, y_j, \widehat{f}_j(\boldsymbol{x}_j)) \tag{10}$$

但し，$\widehat{f}_j(\boldsymbol{x})$ は $\{(\boldsymbol{x}_i, y_i)\}_{i\neq j}$ から学習した結果である．

$P_0(\boldsymbol{x}) = P_1(\boldsymbol{x})$ のとき，LOOCV は期待汎化誤差のほぼ不偏推定量であることが知られている．正確には，LOOCV は $n-1$ 個の標本に対する期待汎化誤差の不偏推定量である [15, 25]．即ち，$n-1$ 個の標本に対する汎化誤差を G' で表せば，

$$\mathop{\mathbb{E}}_{\mathcal{X}_n,\mathcal{Y}_n}\widehat{G}_{LOOCV} = \mathop{\mathbb{E}}_{\mathcal{X}_{n-1},\mathcal{Y}_{n-1}}G' \approx \mathop{\mathbb{E}}_{\mathcal{X}_n,\mathcal{Y}_n}G \tag{11}$$

が成り立つ．しかし，この不偏性は共変量シフト下では成立しない．

この問題に対処するためには，IWERM の時と同様に重要度で重みをつければよい．これを**重要度重み付き CV**（IWCV）と呼ぶ [32]．

$$\widehat{G}_{kIWCV} = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{|\mathcal{T}_j|}\sum_{(\boldsymbol{x},y)\in\mathcal{T}_j}\frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})}\ell(\boldsymbol{x}, y, \widehat{f}_{\mathcal{T}_j}(\boldsymbol{x})) \tag{12}$$

$$\widehat{G}_{LOOIWCV} = \frac{1}{n}\sum_{j=1}^{n}\frac{p_1(\boldsymbol{x}_j)}{p_0(\boldsymbol{x}_j)}\ell(\boldsymbol{x}_j, y_j, \widehat{f}_j(\boldsymbol{x}_j)) \tag{13}$$

LOOIWCV は，共変量シフト下でも期待汎化誤差のほぼ不偏推定量，即ち，$n-1$ 個の標本に対する期待汎化誤差の不偏推定量になっている．IWCV は，任意の損失関数，任意のパラメータ学習法，任意のモデルに対して不偏性を持つため，非常に汎用的である．しかし，IWCV の値を計算するためには繰り返し学習を行なう必要があるため，時間がかかるのが難点である．

IWCV の効果を図 1 の回帰問題の例を用いて確認しよう．図 2-(A) は，図 1 の回帰問題における汎化誤差 G を AIWLS の超パラメータ λ の関数として図示したものである．この例では $\lambda = 0.5$ 付近で最小値を取る．図 2-(E) は 10 重 CV による汎化誤差の推定値である．CV は非常に大きなバイアスを持っており，CV の値を最小にするように λ を決定すると，$\lambda = 0$ が選ばれてしまい，最終的に得られる汎化誤差は非常に大きくなってしまう．図 2-(B) は 10 重IWCV による汎化誤差の推定値である．IWCV を用いるとバイアスはほとんど無くなり，適切な λ を選ぶことができる．

4.2 重要度重み付き部分空間情報量規準

本節では，線形回帰問題に特化した汎化誤差推定法を紹介する．以下，二乗損失に対してモデルが

$$\widehat{f}(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{b} \theta_i \varphi_i(\boldsymbol{x}) \tag{14}$$

で与えられる場合を考える．但し，$\{\varphi_i(\boldsymbol{x})\}_{i=1}^{b}$ は線形独立な基底関数である．このとき，AIWLS の学習結果は

$$\widehat{\boldsymbol{\theta}}_\lambda = \boldsymbol{L}_\lambda \boldsymbol{y} \tag{15}$$

で与えられる．但し，

$$\boldsymbol{L}_\lambda = (\boldsymbol{X}^\top \boldsymbol{D}^\lambda \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{D}^\lambda \tag{16}$$

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top \tag{17}$$

であり，$\boldsymbol{X}$ は (i,j) 要素が $\varphi_j(\boldsymbol{x}_i)$ の $n \times b$ 行列，$\boldsymbol{D}$ は $p_1(\boldsymbol{x}_i)/p_0(\boldsymbol{x}_i)$ を第 i 対角要素とする n 次対角行列である．

この設定のもと，**重要度重み付き部分空間情報量規準**（IWSIC; IW Subspace Information Criterion）は次式で定義される [33].

$$\begin{aligned}\widehat{G}_{IWSIC} =& \langle \boldsymbol{U}\widehat{\boldsymbol{\theta}}_\lambda, \widehat{\boldsymbol{\theta}}_\lambda \rangle - 2\langle \boldsymbol{U}\widehat{\boldsymbol{\theta}}_\lambda, \widehat{\boldsymbol{\theta}}_1 \rangle \\ &+ 2\widehat{\sigma}^2 \mathrm{tr}(\boldsymbol{U}\boldsymbol{L}_\lambda \boldsymbol{L}_1^\top)\end{aligned} \tag{18}$$

但し，$\boldsymbol{U}$ は (i,j) 要素が $\mathbb{E}_{\boldsymbol{t}}\, \varphi_i(\boldsymbol{t})\varphi_j(\boldsymbol{t})$ の b 次元正方行列であり，$\widehat{\boldsymbol{\theta}}_1$ と $\boldsymbol{L}_1$ は $\widehat{\boldsymbol{\theta}}_\lambda$ と $\boldsymbol{L}_\lambda$ において $\lambda = 1$ とおいたものである．また，$\widehat{\sigma}^2 = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}_0\|^2/(n-b)$ である．$\widehat{\boldsymbol{\theta}}_0$ は $\widehat{\boldsymbol{\theta}}_\lambda$ において $\lambda = 0$ とおいたものである．SIC という名前は，元々の規準が部分空間モデルの選択のために提案されたことに由来している [36].

IWSIC は次式を満たす [33].

$$\mathop{\mathbb{E}}_{\mathcal{Y}_n} \widehat{G}_{IWSIC} = \mathop{\mathbb{E}}_{\mathcal{Y}_n} G - C + \mathcal{O}_p(\delta n^{-1/2}) \tag{19}$$

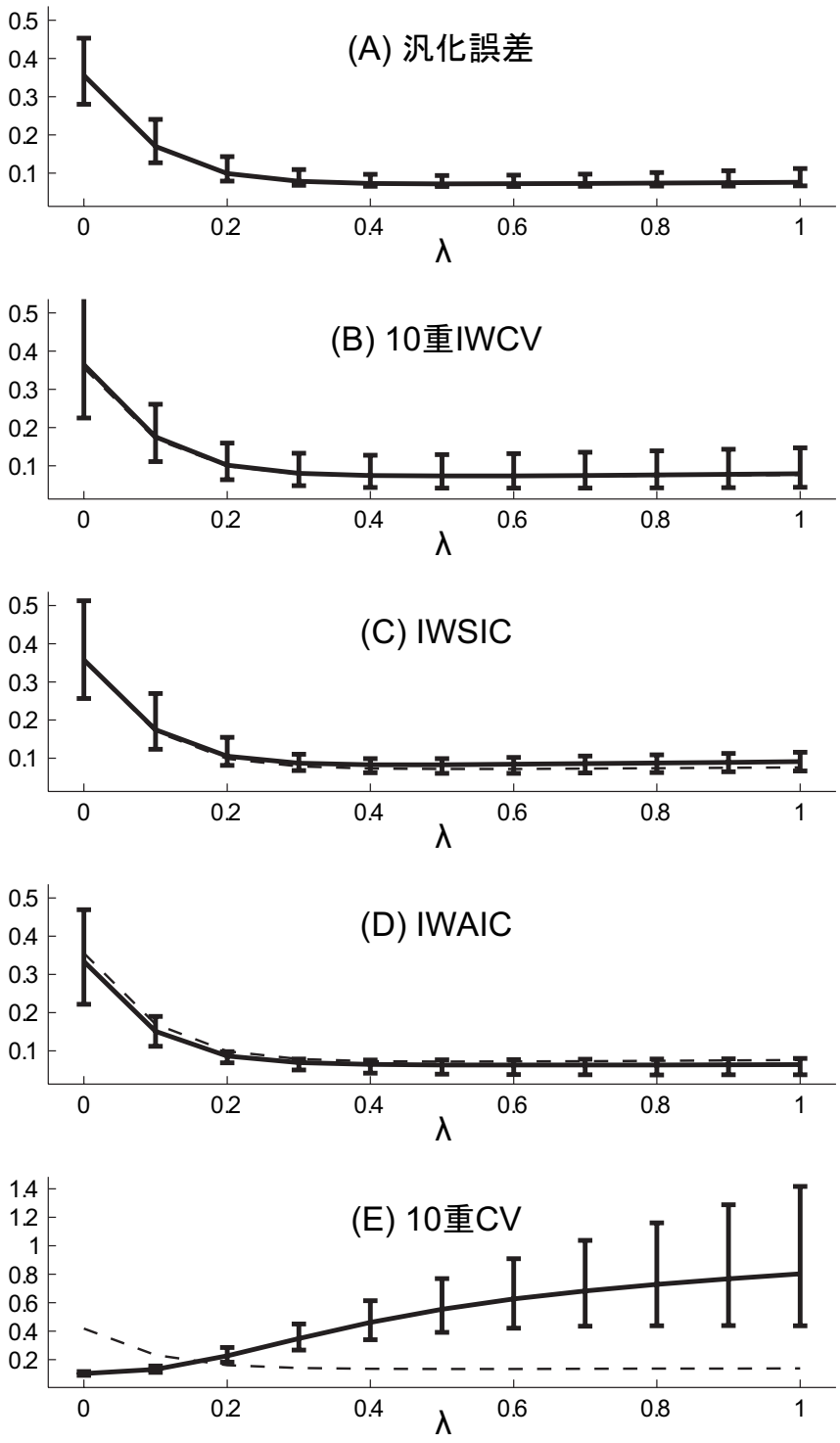

図 2: 汎化誤差とその推定値．グラフの横軸は AIWLS の超パラメータ λ である．下の 4 つのグラフの破線は真の期待汎化誤差である．

但し，C は定数であり，$\delta = \sqrt{\mathbb{E}_{\boldsymbol{u}}\left(\widehat{f}(\boldsymbol{u};\boldsymbol{\theta}^*) - f(\boldsymbol{u})\right)^2}$ はモデルの誤差を表す．式 (19) より，IWSIC は $\mathbb{E}_{\mathcal{Y}_n}$ のもとで期待汎化誤差から定数 C を除いた量の漸近不偏推定量であり，更にその漸近誤差は δ に比例することがわかる．モデルが正しい時，即ち，$\delta = 0$ の時，IWSIC のバイアスは有限標本に対してゼロとなる．

前述した IWCV や後で紹介する IWAIC も，$\mathbb{E}_{\mathcal{Y}_n}$ のもとで漸近誤差が $n^{-1/2}$ のオーダーで減少するという性質を持っている．しかし IWSIC では，漸近誤差が更に δ に比例する．実際に学習を行う場面ではモデルがひどく間違っていることは考えにくいため，δ はそれほど大きくないことが多い．このような場合，訓練標本数が少ない場合でも IWSIC の誤差は小さくなり，実用上非常に有益である．更に，IWSIC は二つの異なるモデルに対する汎化誤差の差を精度良く推定できることが示されている [33]．これは，モデルの良さを比較するモデル選択において重要な性質である．

共変量シフト下での回帰問題に対しては，赤池の情報量規準（AIC; Akaike's Information Criterion）[1] を一般化した**重要度重み付き AIC**（IWAIC）[28] も利用できる．IWAIC は，$\mathbb{E}_{\mathcal{X}_n,\mathcal{Y}_n}$ のもとでは IWSIC よりも収束が速いという性質を持っている．しかし IWSIC と違い，モデルが正しくても IWAIC は有限標本に対して厳密に不偏にはならない．

汎化誤差推定の研究では，$\mathbb{E}_{\mathcal{X}_n,\mathcal{Y}_n}$ のもとでの近似精度を議論することが多い．しかし，現実のモデル選択の場面では訓練標本 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ は一度しか得られない．従って，できる限り訓練標本に関して期待値を取らずに汎化誤差の近似精度を評価することが望ましい．訓練出力 $\{y_i\}_{i=1}^n$ に含まれる雑音は未知なので，訓練出力に関しては平均的に評価せざるを得ないが，訓練入力 $\{\boldsymbol{x}_i\}_{i=1}^n$ はそのものが与えられているため，条件付きで評価することが原理的に可能である．IWSIC ではその立場を前面に押し出している．

図 2-(C) と (D) に図 1 の回帰問題における IWSIC と IWAIC による汎化誤差の推定値を示す．これより，IWSIC と IWAIC によって汎化誤差をほぼ不偏に推定できていることが分かる．また，これらは IWCV と比べて分散が小さい．回帰問題に特化することにより，より精度の良い汎化誤差推定量が得られたのであろう．入力が高次元の複雑な問題に対しては，IWAIC よりも IWSIC の方が推定精度が良いようである [33, 32]．

5 能動学習法

モデル選択と並ぶ教師付き学習の重要な研究課題は，**能動学習**である．能動学習とは，汎化誤差ができるだけ小さくなるように訓練入力 $\{\boldsymbol{x}_i\}_{i=1}^n$，または訓練入力を生成する確率分布 $P_0(\boldsymbol{x})$ を設計する問題である [7, 21]．汎化誤差は未知のため，モデル選択の時と同様に汎化誤差をどれだけ精度良く推定できるかが能動学習研究の本質である．但し，モデル選択の時と異なり，能動学習では訓練出力 $\{y_i\}_{i=1}^n$ を採取する前に汎化誤差を推定する必要があるため，汎化誤差推定はより困難である．以下では，第 4.2 節と同じ線形回帰問題を考える．

線形回帰問題では，汎化誤差の訓練出力 $\{y_i\}_{i=1}^n$ に関する期待値は次のように分解で

きる．

$$\mathop{\mathbb{E}}_{\mathcal{Y}_n} G = B + V + \delta^2 \tag{20}$$

但し，B と V はバイアス（の二乗）と分散であり，それぞれ

$$B = \langle \boldsymbol{U}(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}^*), \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \rangle \tag{21}$$

$$V = \mathop{\mathbb{E}}_{\mathcal{Y}_n} \langle \boldsymbol{U}(\widehat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}} \rangle \tag{22}$$

と表される．但し，$\overline{\boldsymbol{\theta}} = \mathbb{E}_{\mathcal{Y}_n} \widehat{\boldsymbol{\theta}}$ である．

5.1 モデルが正しい場合の能動学習法

古典的な能動学習研究では，学習したい真の関数 $f(\boldsymbol{x})$ がモデル $\widehat{f}(\boldsymbol{x};\boldsymbol{\theta})$ に含まれていることを仮定し，パラメータ学習法として OLS を用いる [7, 21]．モデルが正しいという仮定のもとでは，共変量シフトを意識することなく能動学習を行なうことができる．即ち，$\mathrm{rank}\,(\boldsymbol{X}) = b$ の時バイアス B はゼロとなり，定数 δ^2 を無視すれば，式 (20) の期待汎化誤差は

$$J_{OLS} = \mathrm{tr}(\boldsymbol{U}\boldsymbol{L}_0\boldsymbol{L}_0^\top) = \mathrm{tr}(\boldsymbol{U}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}) \tag{23}$$

に比例する．但し，$\boldsymbol{L}_0$ は $\boldsymbol{L}_\lambda$ において $\lambda = 1$ とおいたものである．この J_{OLS} を最小にするように訓練入力 $\{\boldsymbol{x}_i\}_{i=1}^n$ を最適化すれば汎化誤差を小さくすることができる．この規準は **Q 最適規準**とも呼ばれる [7, 9]．

三角多項式モデルに対しては，n 点の訓練入力の最適な場所を解析的に求めることができる [7, 35]．しかし，一般には n 点の訓練入力を同時に最適化することは困難である．この問題を回避するため，訓練入力を生成する確率分布 $P_0(\boldsymbol{x})$ を最適化するか [31]，あるいは逐次的に訓練入力を決定する [7, 9] ことが多い．

J_{OLS} の前提条件である $\mathrm{rank}\,(\boldsymbol{X}) = b$ を満たすためには，少なくとも b 個の訓練標本が必要である．逐次的に能動学習を行なうとき，通常はランダムに b 個の初期訓練入力を生成し，$b+1$ 番目の訓練入力から最適化する [9]．しかしこの時，初期訓練入力の取り方も分散を最小にするように設計することにより，更に汎化誤差を小さくできることが知られている [34]．

5.2 モデルが正しくない場合の能動学習法

J_{OLS} は簡便であり汎化誤差の改善効果も大きいが，これまで実問題に対してはあまり適用されてこなかったようである．それは，モデルが真の関数を含むという条件が現実的には満たされないことが多いためであろう．この問題を解決すべく，モデルが正しくない場合における能動学習の研究が近年盛んに行なわれている．

モデルが正しくない場合，共変量シフトのため OLS のバイアスは漸近的にもゼロにならない．従って，式 (23) を用いて能動学習を行なうと，分散は小さくなるがバイアスが大きくなり，結果として汎化誤差はあまり小さくならない [31]．一方，IWLS を用いればバイアスは漸近的にゼロになる．そこで，IWLS の分散を小さくするように訓練入力分布 $P_0(\boldsymbol{x})$ を最適化する方法が提案された（ALICE; Active Learning using IWLS based on Conditional Expectation of generalization error）[31]：

$$J_{ALICE} = \mathrm{tr}(\boldsymbol{U}\boldsymbol{L}_1\boldsymbol{L}_1^\top) \tag{24}$$

ALICE では，訓練出力に関する期待値 $\mathbb{E}_{\mathcal{Y}_n}$ のもとでの分散を最小にするように訓練入力分布を最適化している．一方，同様な設定で $\mathbb{E}_{\mathcal{X}_n,\mathcal{Y}_n}$ のもとでの分散を最小にする能動学習法が Wiens によって提案されている [42]．この方法は，最適な確率分布の形状が解析的に導出できるという特徴を持つ．しかし，訓練入力の取り方に関しても期待値を取ると汎化誤差の近似精度が低下してしまうことが示されている [31]．

ALICE を用いれば，モデルが正しくない場合でも効果的に能動学習を行なうことができる．しかし，この手法は理論的には少し問題がある．それは，IWLS のバイアスは漸近的にしかゼロにならず，有限標本に対しては式 (24) で評価している分散と同じオーダーのバイアスが残っているということである．即ち，ALICE では汎化誤差の一部だけを最小化していることになる．

この問題に対処すべく，二段階で能動学習を行なう方法が提案されている [13]．この二段階法では，まず第一段階でいくつかの訓練入力をテスト入力分布に従って生成し，訓練出力を採取する．そして第二段階では，第一段階で集めた訓練標本を使って IWLS の汎化誤差を推定し，それを元に残りの訓練入力を最適化する．この方法では分散だけでなくバイアスもきちんと評価するため，漸近的に最適な能動学習結果が得られる．しかし，バイアスの評価のために一部の訓練標本が使われてしまうため，ユーザが自由に設計できる訓練入力数は実効的に少なくなってしまう．そのため，標本数が有限の現実的な場面では汎化誤差の改善効果はそれほど大きくないようである [31]．

図 3 に，ベンチマークデータに対する能動学習結果を示す（詳細は [31] を参照せよ）．グラフでは，受動学習（テスト入力分布に従って訓練入力を生成）による汎化誤差が 1 になるように値を正規化してある．この結果から，J_{OLS} は不安定であり場合によっては受動学習よりも汎化誤差が大きくなってしまうことがわかる．これは，モデルがほぼ正しい時には J_{OLS} は優れた性能を示すが，そうでない場合は性能が大きく低下してしまうことを示している．一方，ALICE は安定して受動学習を上回っており，他の能動学習法よりも汎化誤差低減効果が大きい．

6 まとめと今後の展望

本稿では，訓練時とテスト時で入力分布が変化する共変量シフトというパラダイムのもと，近年提案されたパラメータ学習法，モデル選択法，能動学習法を紹介した．最近，ブ

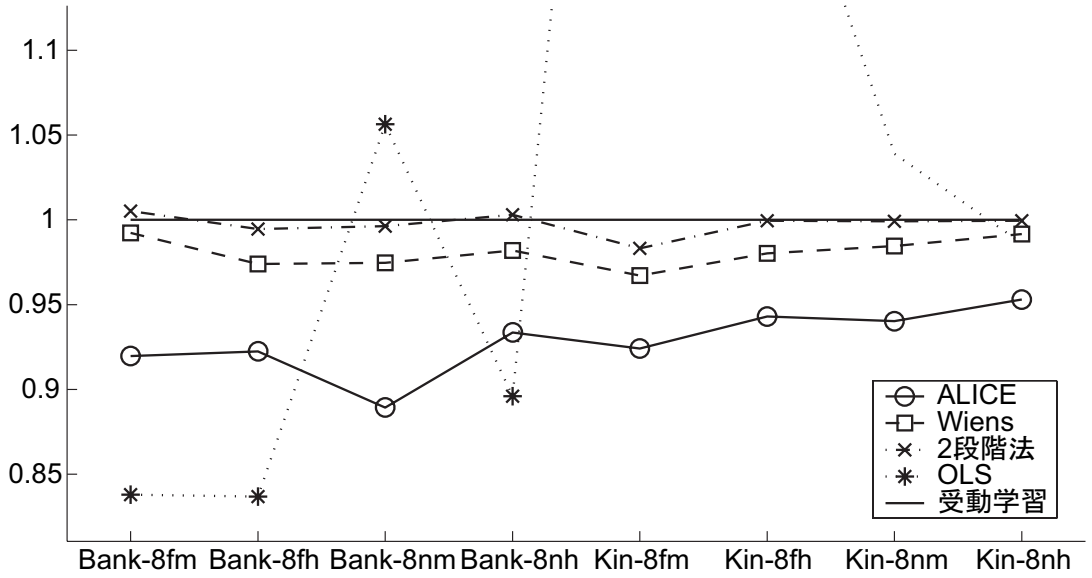

図 3: 各能動学習法による平均汎化誤差.

レイン・コンピュータインターフェースにおける分類問題で，脳の非定常性を共変量シフトとして捕らえ，それを明示的に考慮することにより認識精度が向上することが示された [32]．共変量シフトは強化学習やバイオインフォマティックスなど近年盛んに研究されている応用分野でも頻繁に起こる．今後は共変量シフト下での一般的な学習の方法論のみならず，各応用場面に特化した学習法も開発していく必要がある．

これまで，モデル選択と能動学習の問題は別々に議論されてきた．しかし，これらは両方とも汎化誤差の低減に大きく貢献するため，モデル選択と能動学習を**同時に**行いたいと考えるのは自然なことであろう．しかし，モデルと訓練入力の同時最適化は一筋縄ではいかない．なぜならば，従来のモデル選択法ではあらかじめ固定した訓練入力（と訓練出力）が必要であり，また，従来の能動学習法ではモデルを固定しておく必要があるからである．これを**モデル選択と能動学習のジレンマ**と呼ぶ [37]．これまでに，三角多項式モデル集合に対してこのジレンマをうまく回避する方法が提案されている [37]．しかし，一般にはモデル選択と能動学習を同時に実行することは非常に困難である．モデルと訓練入力を逐次的に選んでいくオンライン型のアルゴリズム [16] を用いることも考えられるが，オンラインアルゴリズムではしばしば良い結果が得られない．なぜならば，最適な訓練入力の場所はモデルに大きく依存し，最適なモデルは標本数によって大きく異なるからである．即ち，オンライン学習過程の初期に選ばれた訓練入力が，最終的に選ばれるモデルにとって好ましいものであるとは限らないのである [24]．一般的な設定のもとでモデル選択と能動学習を同時に行なう手法を開発することは，理論的な観点からも実用的な観点からも非常に重要である．

本稿では，$p_1(\boldsymbol{x})$ をテスト入力の確率密度関数としたが，これをユーザが興味のある入

力領域を示す重み関数と解釈することもできる [28]．この重み関数を局所領域に設定すれば，共変量シフト下での学習は局所学習法 [3, 39] の一つとみなすことができる．このような立場から共変量シフト下での学習法の性質を論じることは非常に興味深い．

教師付き学習の分野には，本稿で扱った頻度主義的な学習法以外にベイズ主義的な学習法があり，盛んに研究されている [16, 41, 22]．ベイズの立場から共変量シフトを論じることも今後の重要な課題であろう [30, 44]．

共変量シフトは，訓練時とテスト時で入出力規則は変わらないが，訓練入力とテスト入力の分布は変化するという状況であった．共変量シフト以外にも，訓練標本とテスト標本の生成規則が異なる状況がある．例えば，分類問題におけるクラス事前確率変化はその一つであろう．これは，訓練時とテスト時でクラスの事前確率 $P(y)$（各クラスの標本の割合）が変化するが，クラスが与えられたもとでの入力分布 $P(\boldsymbol{x}|y)$ は変化しないという状況である [12, 5, 6]．このような状況では，$p_1(\boldsymbol{x})/p_0(\boldsymbol{x})$ の代わりに $p_1(y)/p_0(y)$ で重みをつければ一致性や不偏性が保たれる [14, 32]．これ以外にも訓練標本とテスト標本の生成規則が異なる様々な状況が考えられる [30]．そのような状況を定式化し，各場面にふさわしい学習法を開発することは今後の重要な課題である．

謝辞

本稿の執筆において，Fraunhofer 研究所の Klaus-Robert Müller 氏と川鍋一晃氏，Edinburgh 大学の Amos Storkey 氏と Sethu Vijayakumar 氏，東レエンジニアリング（株）の小川英光氏，東京工業大学の山崎啓介氏，北海道大学の山内康一郎氏との議論を参考にした．本研究の一部は科学研究費補助金若手研究 (B)17700142，基盤研究 (B)18300057，Alexander von Humboldt 財団，EU Erasmus Mundus の援助を受けて行なわれた．

参考文献

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723, 1974.

[2] H. Akaike. Likelihood and the Bayes procedure. In N. J. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics*, pages 141–166, Valencia, 1980. University Press.

[3] C. Atkeson, A. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11:76–113, 1997.

[4] P. Baldi, S. Brunak, and G. A. Stolovitzky. *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, 1998.

[5] N. Chawla, N. Japkowicz, and A. Kolcz, editors. *Proceedings of the ICML2003 Workshop on Learning from Imbalanced Data Sets*, 2003.

[6] N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: Special issue on learning from imbalanced sata sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004.

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

[8] G. S. Fishman. *Monte Carlo: Concepts, Algorithms, and Applications*. Springer-Verlag, Berlin, 1996.

[9] K. Fukumizu. Statistical active learning in multilayer perceptrons. *IEEE Transactions on Neural Networks*, 11(1):17–26, 2000.

[10] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2001.

[11] J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–162, 1979.

[12] N. Japkowicz, editor. *Proceedings of the AAAI2000 Workshop on Learning from Imbalanced Data Sets*, 2000.

[13] T. Kanamori and H. Shimodaira. Active learning algorithm using the maximum weighted log-likelihood estimator. *Journal of Statistical Planning and Inference*, 116(1):149–162, 2003.

[14] Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in non-standard situations. *Machine Learning*, 46(1/3):191–202, 2002.

[15] A. Luntz and V. Brailovsky. On estimation of characters obtained in statistical procedure of recognition. *Technicheskaya Kibernetica*, 3, 1969. in Russian.

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

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

[18] N. Murata, M. Kawanabe, A. Ziehe, K.-R. Müller, and S. Amari. On-line learning in changing environments with applications in supervised and unsupervised learning. *Neural Networks*, 15(4-6):743–760, 2002.

[19] D. Pfeffermann, C. J. Skinner, Holmes D. J., H. Goldstein, and J. Rasbash. Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, series B*, 60(1):23–40, 1998.

[20] J. Platt. A resource-allocating network for function interpolation. *Neural Computation*, 3(2):213–225, 1991.

[21] F. Pukelsheim. *Optimal Design of Experiments*. Wiley, 1993.

[22] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, 2006.

[23] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[24] N. Rubens and M. Sugiyama. Ensemble active learning: Risk hedging in active learning with model selection, 2006. submitted.

[25] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[26] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[27] C. R. Shelton. *Importance Sampling for Reinforcement Learning with Multiple Objectives*. PhD thesis, Massachusetts Institute of Technology, 2001.

[28] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[29] C. J. Skinner, D. Holt, and T. M. F. Smith. *Analysis of Complex Surveys*. Wiley, New York, 1989.

[30] A. Storkey and M. Sugiyama. Mixture regression for covariate shift, 2006. submitted.

[31] M. Sugiyama. Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research*, 7:141–166, Jan. 2006.

[32] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation, 2006. submitted.

[33] M. Sugiyama and K.-R. Müller. Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, 23(4):249–279, 2005.

[34] M. Sugiyama and H. Ogawa. Incremental active learning for optimal generalization. *Neural Computation*, 12(12):2909–2940, 2000.

[35] M. Sugiyama and H. Ogawa. Active learning for optimal generalization in trigonometric polynomial models. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E84-A(9):2319–2329, 2001.

[36] M. Sugiyama and H. Ogawa. Subspace information criterion for model selection. *Neural Computation*, 13(8):1863–1889, 2001.

[37] M. Sugiyama and H. Ogawa. Active learning with model selection—Simultaneous optimization of sample points and models for trigonometric polynomial models. *IEICE Transactions on Information and Systems*, E86-D(12):2753–2763, 2003.

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

[39] S. Vijayakumar, A. D'Souza, and S. Schaal. Incremental online learning in high dimensions. *Neural Computation*, 17(12):2602–2634, December 2005.

[40] G. Wahba. *Spline Model for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia and Pennsylvania, 1990.

[41] S. Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13(4):899–933, 2001.

[42] D. P. Wiens. Robust weights and designs for biased regression models: Least squares and generalized M-estimation. *Journal of Statistical Planning and Inference*, 83(2):395–412, 2000.

[43] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002.

[44] K. Yamazaki, S. Watanabe, M. Sugiyama, K.-R. Müller, and M. Kawanabe. Asymptotic Bayes estimation under the covariate shift, 2006. submitted.