Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Covariate Shift Adaptation by Importance Weighted Cross Validation

Masashi Sugiyama SUGI@CS.TITECH.AC.JP
Department of Computer Science
Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan
and
Institute of Perception, Action, and Behaviour
University of Edinburgh
The King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, UK

Matthias Krauledat MATTHIAS.KRAULEDAT@FIRST.FHG.DE
Department of Computer Science
Technical University Berlin
Franklinstr. 28/29, 10587 Berlin, Germany
and
Fraunhofer FIRST.IDA
Kekuléstr. 7, 12489 Berlin, Germany

Klaus-Robert Müller KLAUS@FIRST.FHG.DE
Department of Computer Science
Technical University Berlin
Franklinstr. 28/29, 10587 Berlin, Germany
and
Fraunhofer FIRST.IDA
Kekuléstr. 7, 12489 Berlin, Germany

Editor: Yoshua Bengio

Abstract

A common assumption in supervised learning is that the input points in the training set follow the *same* probability distribution as the input points that will be given in the future test phase. However, this assumption is not satisfied, for example, when the outside of the training region is extrapolated. The situation where the training input points and test input points follow *different* distributions while the conditional distribution of output values given input points is unchanged is called the *covariate shift*. Under the covariate shift, standard model selection techniques such as cross validation do not work as desired since its unbiasedness is no longer maintained. In this paper, we propose a new method called *importance weighted cross validation* (IWCV), for which we prove its unbiasedness even under the covariate shift. The IWCV procedure is the only one that can be applied for unbiased classification under covariate shift, whereas alternatives to IWCV exist for regression. The usefulness of our proposed method is illustrated by simulations, and furthermore demonstrated in the brain-computer interface, where strong non-stationarity effects can be seen between training and test sessions.

1. Introduction

The goal of supervised learning is to infer an unknown input-output dependency from training samples, by which output values for unseen test input points can be estimated. When developing a method of supervised learning, it is commonly assumed that the input points in the training set and the input points used for testing follow the *same* probability distribution (e.g., Wahba, 1990; Bishop, 1995; Vapnik, 1998; Duda et al., 2001; Hastie et al., 2001; Schölkopf and Smola, 2002). However, this common assumption is not fulfilled, for example, when we extrapolate outside of the training region[1] or when training input points are designed by an active learning (experimental design) algorithm. The situation where the training input points and test input points follow different probability distributions but the conditional distributions of output values given input points are unchanged is called the *covariate shift* (Shimodaira, 2000). For data from many applications such as off-policy reinforcement learning (Shelton, 2001), spam filtering (Bickel and Scheffer, 2007), bioinformatics (Baldi et al., 1998; Borgwardt et al., 2006) or brain-computer interfacing (Wolpaw et al., 2002), the covariate shift phenomenon is conceivable. Sample selection bias (Heckman, 1979) in economics may also include a form of the covariate shift. Illustrative examples of covariate shift situations are depicted in Figures 1 and 3.

In this paper, we develop a new learning method and prove that we can alleviate misestimation due to covariate shift. From the beginning, we note that all the theoretical discussions will be made under the assumption that the *ratio* of test and training input densities at training input points is known; in experimental studies, the density ratio will be replaced by their empirical estimates and the practical performance of our approach will be evaluated.

Model selection is one of the key ingredients in machine learning. However, under the covariate shift, a standard model selection technique such as *cross validation* (CV) (Stone, 1974; Wahba, 1990) does not work as desired; more specifically, the unbiasedness that guarantees the accuracy of CV does not hold under the covariate shift anymore. To cope with this problem, we propose a novel variant of CV called *importance weighted CV* (IWCV). We prove that IWCV gives an almost unbiased estimate of the risk even under the covariate shift. Model selection under the covariate shift has been studied so far only by few researchers (e.g., Shimodaira, 2000; Sugiyama and Müller, 2005)—existing methods have a number of limitations, e.g., in the loss function, parameter learning method, and model. In particular, the existing methods can not be applied to classification scenarios. On the other hand, the proposed IWCV overcomes these limitations: it allows for *any* loss function, parameter learning method, and model; even non-parametric learning methods can be employed. To the best of our knowledge, the proposed IWCV is the first method that can be successfully applied to model selection in covariate-shifted classification tasks. The usefulness of the proposed method is demonstrated in the brain-computer interface applications, in which existing methods for covariate shift compensation could not be employed.

2. Problem Formulation

In this section, we formulate the supervised learning problem with the covariate shift, and review existing learning methods.

1. The term 'extrapolation' could have been defined in a narrow sense as prediction in regions with *no* training samples. On the other hand, the situation we are considering here is 'weak' extrapolation; prediction is carried out in the region where only a small number of training samples is available.

2.1 Supervised Learning under Covariate Shift

Let us consider the supervised learning problem of estimating an unknown input-output dependency from training samples. Let $\mathcal{T} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be the training samples, where $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is an i.i.d. training input point following a probability distribution $P_{train}(\boldsymbol{x})$ and $y_i \in \mathcal{Y} \subset \mathbb{R}$ is a corresponding training output value following a conditional probability distribution $P(y|\boldsymbol{x})$. $P(y|\boldsymbol{x})$ may be regarded as the sum of true output $f(\boldsymbol{x})$ and noise.

Let $\ell(\boldsymbol{x}, y, \widehat{y}) : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ be the loss function, which measures the discrepancy between the true output value y at an input point $\boldsymbol{x}$ and its estimate $\widehat{y}$. Let us employ a parametric model $\widehat{f}(\boldsymbol{x}; \boldsymbol{\theta})$ for estimating the output value y, where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^b$. Note that the range of application of our proposed method given in Section 3 includes non-parametric methods, but we focus on a parametric setting for simplicity. A model $\widehat{f}(\boldsymbol{x}; \boldsymbol{\theta})$ is said to be *correctly specified* if there exists a parameter $\boldsymbol{\theta}^*$ such that $\widehat{f}(\boldsymbol{x}; \boldsymbol{\theta}^*) = f(\boldsymbol{x})$; otherwise the model is said to be *misspecified*. In practice, the model used for learning would be misspecified to a greater or lesser extent. For this reason, we do not assume that the model is correct in this paper. The goal of supervised learning is to determine the value of the parameter $\boldsymbol{\theta}$ so that output values for unlearned test input points are accurately estimated.

Let us consider a test sample, which is not given to the user in the training phase, but will be given in the test phase in the future. We denote the test sample by $(\boldsymbol{t}, u)$, where $\boldsymbol{t} \in \mathcal{X}$ is a test input point and $u \in \mathcal{Y}$ is a corresponding test output value. The test error expected over test samples is expressed as

$$\mathbb{E}_{\boldsymbol{t},u}\left[\ell(\boldsymbol{t}, u, \widehat{f}(\boldsymbol{t}; \widehat{\boldsymbol{\theta}}))\right], \tag{1}$$

where $\mathbb{E}$ denotes the expectation. Note that the learned parameter $\widehat{\boldsymbol{\theta}}$ generally depends on the training set $\mathcal{T} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. In the following, we consider the expected test error over the training samples, which is called the *risk* or the *generalization error*:

$$R^{(n)} \equiv \mathbb{E}_{\{\boldsymbol{x}_i, y_i\}_{i=1}^n, \boldsymbol{t}, u}\left[\ell(\boldsymbol{t}, u, \widehat{f}(\boldsymbol{t}; \widehat{\boldsymbol{\theta}}))\right]. \tag{2}$$

In standard supervised learning theories, the test sample $(\boldsymbol{t}, u)$ is assumed to follow $P(u|\boldsymbol{t})P_{train}(\boldsymbol{t})$, which is the *same* probability distribution as for the training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ (e.g., Wahba, 1990; Bishop, 1995; Vapnik, 1998; Duda et al., 2001; Hastie et al., 2001; Schölkopf and Smola, 2002). On the other hand, in this paper, we consider the situation under the *covariate shift*, i.e., the conditional distribution $P(u|\boldsymbol{t})$ remains unchanged, but the test input point $\boldsymbol{t}$ follows a different probability distribution $P_{test}(\boldsymbol{x})$. Illustrative examples of covariate shift situations are depicted in Figures 1 and 3.

Let $p_{train}(\boldsymbol{x})$ and $p_{test}(\boldsymbol{x})$ be the probability density functions corresponding to the input distributions $P_{train}(\boldsymbol{x})$ and $P_{test}(\boldsymbol{x})$, respectively. In the following theoretical discussions, we assume that the *ratio* of test and training input densities at training input points,

$$\frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)}, \tag{3}$$

is finite and known. We refer to the expression (3) as *importance* à la *importance sampling* (Fishman, 1996). In practical situations where the importance is unknown, we may replace them by empirical estimates (see Sections 4 and 5).

2.2 Empirical Risk Minimization and Its Importance Weighted Variants

A standard method to learn the parameter $\boldsymbol{\theta}$ would be *empirical risk minimization* (ERM) (e.g., Vapnik, 1998; Schölkopf and Smola, 2002):

$$\widehat{\boldsymbol{\theta}}_{ERM} \equiv \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{x}_i, y_i, \widehat{f}(\boldsymbol{x}_i; \boldsymbol{\theta})) \right]. \tag{4}$$

If $P_{train}(\boldsymbol{x}) = P_{test}(\boldsymbol{x})$, ERM provides a *consistent* estimator[2] (Shimodaira, 2000). However, under the covariate shift where $P_{train}(\boldsymbol{x}) \neq P_{test}(\boldsymbol{x})$, ERM is not generally consistent anymore[3] (Shimodaira, 2000):

$$\lim_{n\to\infty} \left(\widehat{\boldsymbol{\theta}}_{ERM} \right) \neq \boldsymbol{\theta}^*, \tag{5}$$

where

$$\boldsymbol{\theta}^* \equiv \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \left(\mathbb{E}_{\boldsymbol{t},u} \left[\ell(\boldsymbol{t}, u, \widehat{f}(\boldsymbol{t}; \boldsymbol{\theta})) \right] \right). \tag{6}$$

Under the covariate shift, the following *importance weighted ERM* (IWERM) is consistent (Shimodaira, 2000):

$$\widehat{\boldsymbol{\theta}}_{IWERM} \equiv \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)} \ell(\boldsymbol{x}_i, y_i, \widehat{f}(\boldsymbol{x}_i; \boldsymbol{\theta})) \right], \tag{7}$$

which satisfies even for a misspecified model

$$\lim_{n\to\infty} \left(\widehat{\boldsymbol{\theta}}_{IWERM} \right) = \boldsymbol{\theta}^*. \tag{8}$$

Although IWERM is consistent, it is not *efficient* and can be rather unstable (Shimodaira, 2000). Therefore, IWERM may not be optimal in practical finite sample cases; a slightly stabilized variant of IWERM could be practically better than plain IWERM. The trade-off between consistency and stability could be controlled, for example, by weakening the weight (*Adaptive IWERM*; AIWERM) or by adding a regularizer to the empirical risk (*Regularized IWERM*; RIWERM):

$$\widehat{\boldsymbol{\theta}}_{AIWERM} \equiv \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)} \right)^{\lambda} \ell(\boldsymbol{x}_i, y_i, \widehat{f}(\boldsymbol{x}_i; \boldsymbol{\theta})) \right], \tag{9}$$

$$\widehat{\boldsymbol{\theta}}_{RIWERM} \equiv \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)} \ell(\boldsymbol{x}_i, y_i, \widehat{f}(\boldsymbol{x}_i; \boldsymbol{\theta})) + \gamma R(\boldsymbol{\theta}) \right], \tag{10}$$

where $0 \leq \lambda \leq 1$, $\gamma \geq 0$, and $R(\boldsymbol{\theta})$ is some regularization functional.

The above AIWERM and RIWERM methods are just examples; there may be many other possibilities of controlling the trade-off between consistency and stability. We note that the methodology we propose in this paper is valid for *any* parameter learning method; this means that, e.g., an importance weighted variant of support vector machines (Vapnik, 1998; Schölkopf and Smola, 2002; Huang et al., 2007) or graph regularization techniques (Bousquet et al., 2004; Belkin and Niyogi, 2004; Hein, 2006) can also be employed.

2. For a correctly specified model, an estimator is said to be consistent if it converges in probability to the *true* parameter. For a misspecified model, we say that an estimator is consistent if it converges in probability to the *optimal* parameter in the model (i.e., the optimal approximation of the learning target under the risk (2) within the model).
3. If the model is correct, ERM is still consistent even under the covariate shift.

2.3 Cross Validation Estimate of Risk

The value of the tuning parameter, say λ in Eq.(9), controls the trade-off between the consistency and stability. Therefore, we need to perform *model selection* for determining the value of λ. *Cross validation* (CV) is a popular method for model selection (Stone, 1974; Wahba, 1990). A basic idea of CV is to divide the training set into 'training part' and 'validation part'—a learning machine is trained using the training part and is tested using the validation part, by which the risk is estimated. Below, we give a slightly more formal description of the CV procedure, which will be used in the next section.

Let us randomly divide the training set $\mathcal{T} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ into k disjoint non-empty subsets $\{\mathcal{T}_i\}_{i=1}^k$. Let $\widehat{f}_{\mathcal{T}_j}(\boldsymbol{x})$ be a function learned from $\{\mathcal{T}_i\}_{i\neq j}$. Then the *$k$-fold CV* ($k$CV) estimate of the risk $R^{(n)}$ is given by

$$\widehat{R}^{(n)}_{kCV} \equiv \frac{1}{k}\sum_{j=1}^{k}\frac{1}{|\mathcal{T}_j|}\sum_{(\boldsymbol{x},y)\in\mathcal{T}_j}\ell(\boldsymbol{x}, y, \widehat{f}_{\mathcal{T}_j}(\boldsymbol{x})), \tag{11}$$

where $|\mathcal{T}_j|$ is the number of samples in the subset $\mathcal{T}_j$. When $k = n$, kCV is particularly called *leave-one-out cross validation* (LOOCV).

$$\widehat{R}^{(n)}_{LOOCV} \equiv \frac{1}{n}\sum_{j=1}^{n}\ell(\boldsymbol{x}_j, y_j, \widehat{f}_j(\boldsymbol{x}_j)), \tag{12}$$

where $\widehat{f}_j(\boldsymbol{x})$ is a function learned from $\{(\boldsymbol{x}_i, y_i)\}_{i\neq j}$.

It is known that, if $P_{train}(\boldsymbol{x}) = P_{test}(\boldsymbol{x})$, LOOCV gives an *almost* unbiased estimate of the risk; more precisely, LOOCV gives an unbiased estimate of the risk with $n-1$ samples (Luntz and Brailovsky, 1969; Schölkopf and Smola, 2002):

$$\mathbb{E}_{\{\boldsymbol{x}_i, y_i\}_{i=1}^n}\left[\widehat{R}^{(n)}_{LOOCV}\right] = R^{(n-1)} \approx R^{(n)}. \tag{13}$$

However, this useful property is no longer true under the covariate shift. In the following section, we give a novel modified cross validation method which still maintains the 'almost unbiasedness' property even under the covariate shift.

3. Importance Weighted Cross Validation

To compensate for the effect of the covariate shift in the CV procedure, we propose the following *importance weighted CV* (IWCV):

$$\widehat{R}^{(n)}_{kIWCV} \equiv \frac{1}{k}\sum_{j=1}^{k}\frac{1}{|\mathcal{T}_j|}\sum_{(\boldsymbol{x},y)\in\mathcal{T}_j}\frac{p_{test}(\boldsymbol{x})}{p_{train}(\boldsymbol{x})}\ell(\boldsymbol{x}, y, \widehat{f}_{\mathcal{T}_j}(\boldsymbol{x})), \tag{14}$$

or

$$\widehat{R}^{(n)}_{LOOIWCV} \equiv \frac{1}{n}\sum_{j=1}^{n}\frac{p_{test}(\boldsymbol{x}_j)}{p_{train}(\boldsymbol{x}_j)}\ell(\boldsymbol{x}_j, y_j, \widehat{f}_j(\boldsymbol{x}_j)). \tag{15}$$

That is, the validation error in the CV procedure is weighted according to the importance.

The following lemma shows that LOOIWCV gives an almost unbiased estimate of the risk even under the covariate shift.

Lemma 1

$$\mathbb{E}_{\{\boldsymbol{x}_i,y_i\}_{i=1}^n}\left[\widehat{R}^{(n)}_{LOOIWCV}\right] = R^{(n-1)}. \tag{16}$$

Proof Since the conditional distribution $P(y|\boldsymbol{x})$ does not change between the training and test phases, we have, for any $j \in \{1,\ldots,n\}$,

$$\begin{aligned}
&\mathbb{E}_{\{\boldsymbol{x}_i,y_i\}_{i=1}^n}\left[\frac{p_{test}(\boldsymbol{x}_j)}{p_{train}(\boldsymbol{x}_j)}\ell(\boldsymbol{x}_j,y_j,\widehat{f}_j(\boldsymbol{x}_j))\right]\\
&\quad= \mathbb{E}_{\{\boldsymbol{x}_i,y_i\}_{i\neq j},\boldsymbol{x}_j,y_j}\left[\frac{p_{test}(\boldsymbol{x}_j)}{p_{train}(\boldsymbol{x}_j)}\ell(\boldsymbol{x}_j,y_j,\widehat{f}_j(\boldsymbol{x}_j))\right]\\
&\quad= \mathbb{E}_{\{\boldsymbol{x}_i,y_i\}_{i\neq j},y_j}\left[\int_{\mathcal{X}}\frac{p_{test}(\boldsymbol{x}_j)}{p_{train}(\boldsymbol{x}_j)}\ell(\boldsymbol{x}_j,y_j,\widehat{f}_j(\boldsymbol{x}_j))p_{train}(\boldsymbol{x}_j)d\boldsymbol{x}_j\right]\\
&\quad= \mathbb{E}_{\{\boldsymbol{x}_i,y_i\}_{i\neq j},y_j}\left[\int_{\mathcal{X}}p_{test}(\boldsymbol{x}_j)\ell(\boldsymbol{x}_j,y_j,\widehat{f}_j(\boldsymbol{x}_j))d\boldsymbol{x}_j\right]\\
&\quad= \mathbb{E}_{\{\boldsymbol{x}_i,y_i\}_{i\neq j},u}\left[\int_{\mathcal{X}}\ell(\boldsymbol{t},u,\widehat{f}_j(\boldsymbol{t}))p_{test}(\boldsymbol{t})d\boldsymbol{t}\right]\\
&\quad= \mathbb{E}_{\{\boldsymbol{x}_i,y_i\}_{i\neq j},\boldsymbol{t},u}\left[\ell(\boldsymbol{t},u,\widehat{f}_j(\boldsymbol{t}))\right]\\
&\quad= R^{(n-1)}.
\end{aligned} \tag{17}$$

Then we have

$$\begin{aligned}
\mathbb{E}_{\{\boldsymbol{x}_i,y_i\}_{i=1}^n}\left[\widehat{R}^{(n)}_{LOOIWCV}\right] &= \mathbb{E}_{\{\boldsymbol{x}_i,y_i\}_{i=1}^n}\left[\frac{1}{n}\sum_{j=1}^n\frac{p_{test}(\boldsymbol{x}_j)}{p_{train}(\boldsymbol{x}_j)}\ell(\boldsymbol{x}_j,y_j,\widehat{f}_j(\boldsymbol{x}_j))\right]\\
&= \frac{1}{n}\sum_{j=1}^n\mathbb{E}_{\{\boldsymbol{x}_i,y_i\}_{i=1}^n}\left[\frac{p_{test}(\boldsymbol{x}_j)}{p_{train}(\boldsymbol{x}_j)}\ell(\boldsymbol{x}_j,y_j,\widehat{f}_j(\boldsymbol{x}_j))\right]\\
&= \frac{1}{n}\sum_{j=1}^n R^{(n-1)}\\
&= R^{(n-1)},
\end{aligned} \tag{18}$$

which concludes the proof. ■

As proved above, the simple variant of LOOCV called LOOIWCV provides an unbiased estimate of the risk with $n-1$ samples even under the covariate shift. A similar proof is also possible for kIWCV, although its bias may be larger than LOOIWCV. Note that we did not assume any condition on the loss function in the above proof. This means that the almost unbiasedness is valid for any loss function including non-smooth losses such as the 0/1-loss. Also, we did not make a model-specific assumption in the proof. Therefore, the almost unbiasedness holds for any model; even non-identifiable models (Watanabe, 2001) such as multi-layer perceptrons or Gaussian mixture models are included. Furthermore, the above proof does not depend on the method of parameter learning. This means that the almost unbiasedness is valid for any parameter learning method; even non-parametric learning methods are allowed.

4. Numerical Examples

In this section, we illustrate how IWCV works using toy regression and classification data sets. More simulation results can be found in a separate technical report (Sugiyama et al., 2006).

4.1 Toy Regression Problem

Here we illustrate the behavior of the proposed and existing generalization error estimators using a simple one-dimensional regression data set.

We use the following linear model for learning:

$$\widehat{f}(\boldsymbol{x};\boldsymbol{\theta}) = \theta_0 + \sum_{i=1}^{d} \theta_i x^{(i)}, \tag{19}$$

where $x^{(i)}$ is the ith element of $\boldsymbol{x}$ and d is the input dimensionality. The parameter vector $\boldsymbol{\theta}$ is learned by *adaptive importance weighted least-squares* (AIWLS):

$$\widehat{\boldsymbol{\theta}}_{AIWLS} \equiv \operatorname*{argmin}_{\boldsymbol{\theta}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)} \right)^{\lambda} \left(\widehat{f}(\boldsymbol{x}_i;\boldsymbol{\theta}) - y_i \right)^2 \right], \tag{20}$$

where $0 \leq \lambda \leq 1$; λ is chosen later by a model selection method. For the linear model (19), the above minimizer $\widehat{\boldsymbol{\theta}}_{AIWLS}$ is given analytically by

$$\widehat{\boldsymbol{\theta}}_{AIWLS} = (\boldsymbol{X}^\top \boldsymbol{D}^\lambda \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{D}^\lambda \boldsymbol{y}, \tag{21}$$

where

$$\boldsymbol{X} \equiv \begin{pmatrix} 1 & \boldsymbol{x}_1^\top \\ 1 & \boldsymbol{x}_2^\top \\ & \vdots \\ 1 & \boldsymbol{x}_n^\top \end{pmatrix}, \tag{22}$$

which is assumed to have rank $d+1$; $\boldsymbol{D}$ is the diagonal matrix with the ith diagonal element $D_{i,i} = p_{test}(\boldsymbol{x}_i)/p_{train}(\boldsymbol{x}_i)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$.

Let the learning target function be $f(x) = \mathrm{sinc}(x)$ and let the training and test input densities be

$$p_{train}(x) = \phi(x; 1, (1/2)^2), \tag{23}$$

$$p_{test}(x) = \phi(x; 2, (1/4)^2), \tag{24}$$

where $\phi(x; \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2. This setting implies that we are considering an extrapolation problem (see Figure 1(A)). We create the training output value y_i ($i = 1, 2, \ldots, n$) as $y_i = f(x_i) + \epsilon_i$, where $\{\epsilon_i\}_{i=1}^n$ have density $\phi(\epsilon; 0, (1/4)^2)$. Let the number of training samples be $n = 150$.

Figure 1 (B)–(D) illustrate the true function, a realization of training samples, learned functions by AIWLS (20) with $\lambda = 0, 0.5, 1$, and a realization of test samples. For this particular realization, $\lambda = 0.5$ appears to work very well. However, the best choice of λ depends on the realization of samples and λ needs to be carefully chosen by a model selection method. We randomly create

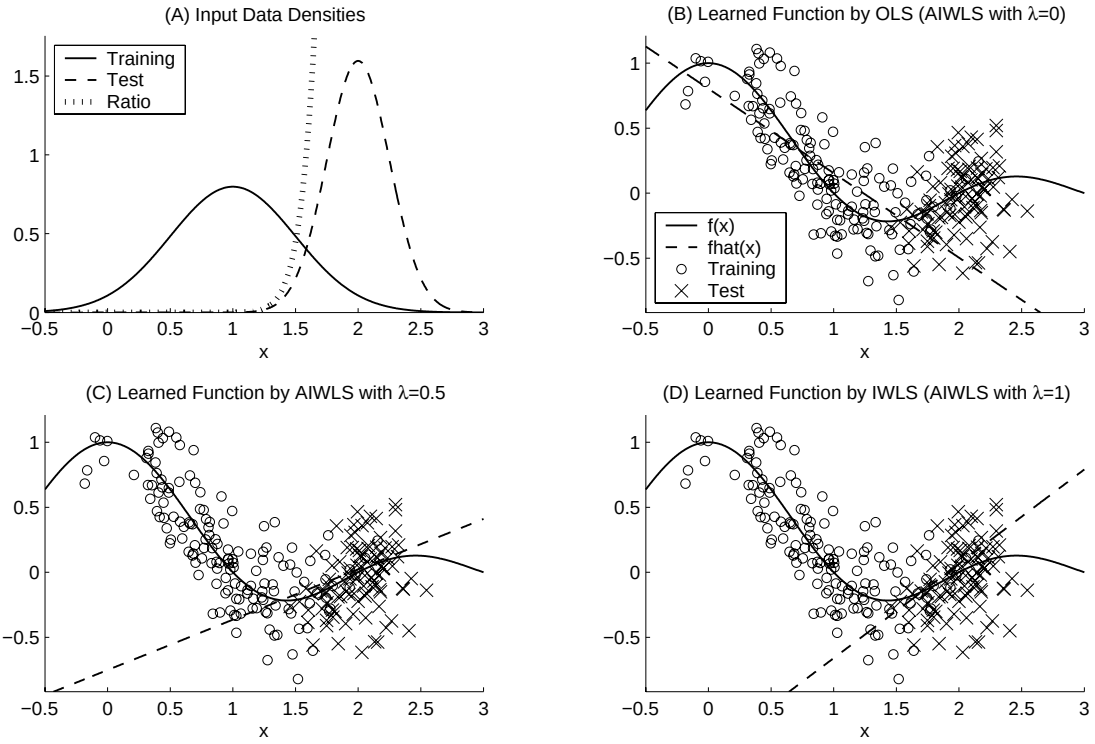


Figure 1: An illustrative regression example of extrapolation by fitting a linear function. (A) The probability density functions of the training and test input points and their ratio. (B)–(D) The learning target function $f(x)$ (the solid line), training samples ('$\circ$'), a learned function $\widehat{f}(x)$ (the dashed line), and test samples ('$\times$'). Note that the test samples are not given in the training phase; they are plotted in the graph for illustration purposes.

$\{x_i, \epsilon_i\}_{i=1}^n$ and calculate the scores of 10-fold IWCV and 10-fold CV for $\lambda = 0, 0.1, 0.2, \ldots, 1$. This procedure is repeated 1000 times.

Top graphs in Figure 2 depicts the mean and standard deviation of the test error and its estimate by each method, as functions of the tuning parameter λ in AIWLS (20). Note that the mean of the test error corresponds to the true risk (see Eqs.(1) and (2)). The graphs show that IWCV gives reasonably good unbiased estimates of the risk, while CV is heavily biased.

Next we investigate the model selection performance: λ is chosen from $\{0, 0.1, 0.2, \ldots, 1\}$ so that the score of each method is minimized. Bottom graphs in Figure 2 depict the histogram of the minimizer of each score. The mean and standard deviation of the test error when λ is chosen by each method are described below the graphs. The numbers show that IWCV gives much smaller test errors than CV (the difference is significant by the *t-test* at the significance level 1%).

We also carried out the same simulation under unknown training and test densities; they are estimated by maximum likelihood fitting of a single Gaussian model or a Gaussian kernel density estimator with variance determined by *Silverman's rule-of-thumb bandwidth selection rule* (Silverman, 1986; Härdle et al., 2004). For estimating the test input density, we draw 100 unlabeled samples following $P_{test}(\boldsymbol{x})$. The simulation results had very similar trends to the case with known densities (therefore we omit the detail), although the error gets slightly larger. This implies that, for

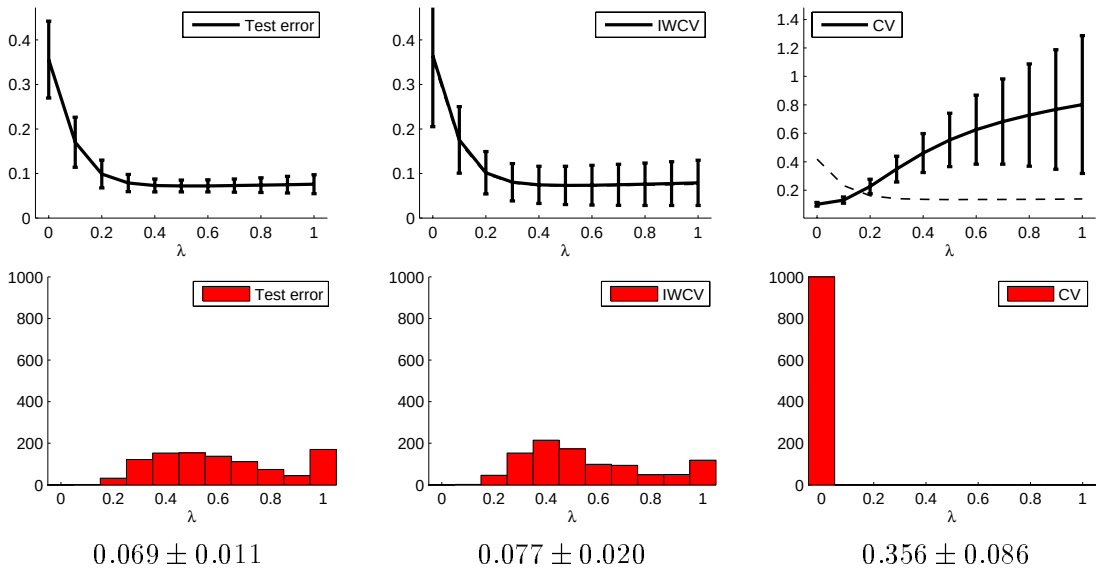


Figure 2: Top graphs: Test error and its estimates as functions of the tuning parameter λ in AIWLS (20). Dashed curves in the middle and right graphs depict the mean test error (i.e., the mean of the left graph) for clear comparison. Bottom graphs: Histograms of the minimizer. The numbers below the graphs are the means and standard deviations of the test error when λ is selected by each method.

this toy regression problem, estimating the densities from data does not significantly degrade the quality of learning.

The above simulation results illustrate that IWCV performs quite well in covariate-shifted regression tasks.

4.2 Toy Classification Problem

Through the above regression examples, we found that IWCV works quite well. Here, we apply IWCV to a toy classification problem where the 0/1-loss is used for computing the test error.

Let us consider a binary classification problem on the two-dimensional input space. We define the class posterior probabilities given input $\boldsymbol{x}$ by

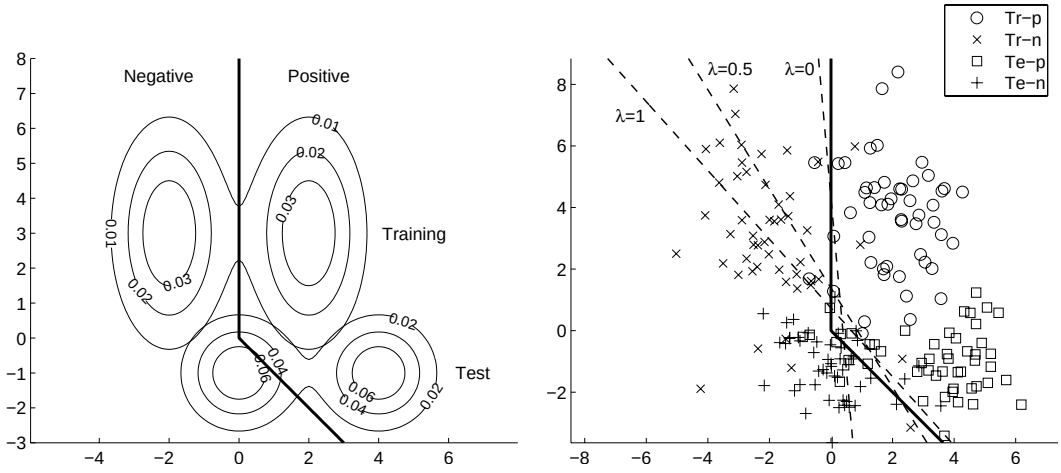
$$p(y=+1|\boldsymbol{x}) = \frac{1+\tanh\left(x^{(1)}+\min(0,x^{(2)})\right)}{2}, \tag{25}$$

where $\boldsymbol{x} = (x^{(1)}, x^{(2)})^\top$ and $p(y=-1|\boldsymbol{x}) = 1 - p(y=+1|\boldsymbol{x})$. The optimal decision boundary, i.e., a set of all $\boldsymbol{x}$ such that $p(y=+1|\boldsymbol{x}) = p(y=-1|\boldsymbol{x})$, is illustrated in Figure 3(b).

Let the training and test input densities be

$$p_{train}(\boldsymbol{x}) = \frac{1}{2}\phi\left(\boldsymbol{x}; \begin{pmatrix}-2\\3\end{pmatrix}, \begin{pmatrix}1&0\\0&2\end{pmatrix}\right) + \frac{1}{2}\phi\left(\boldsymbol{x}; \begin{pmatrix}2\\3\end{pmatrix}, \begin{pmatrix}1&0\\0&2\end{pmatrix}\right), \tag{26}$$

$$p_{test}(\boldsymbol{x}) = \frac{1}{2}\phi\left(\boldsymbol{x}; \begin{pmatrix}0\\-1\end{pmatrix}, \begin{pmatrix}1&0\\0&1\end{pmatrix}\right) + \frac{1}{2}\phi\left(\boldsymbol{x}; \begin{pmatrix}4\\-1\end{pmatrix}, \begin{pmatrix}1&0\\0&1\end{pmatrix}\right), \tag{27}$$



(a) Contours of training and test input densities.

(b) Optimal decision boundary (solid line) and learned boundaries (dashed lines). '$\circ$' and '$\times$' denote the positive and negative training samples, while '$\square$' and '$+$' denote the positive and negative test samples. Note that the test samples are not given in the training phase; they are plotted in the figure for illustration purposes.

Figure 3: Toy classification problem.

where $\phi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. This setting implies that we are considering an extrapolation problem. Contours of the training and test input densities are illustrated in Figure 3(a).

Let $n = 500$ and we create training input points $\{\boldsymbol{x}_i\}_{i=1}^n$ following $P_{train}(\boldsymbol{x})$ and training output labels $\{y_i\}_{i=1}^n$ following $P(y|\boldsymbol{x}_i)$. Similarly, we create 500 test input points $\{\boldsymbol{t}_i\}_{i=1}^{500}$ following $P_{test}(\boldsymbol{x})$ and test output labels $\{u_i\}_{i=1}^{500}$ following $P(u|\boldsymbol{t}_i)$.

We use the linear model (19) combined with AIWLS (20) for learning. The classification result $\widehat{u}$ of a test sample $\boldsymbol{t}$ is obtained by the sign of the output of the learned function:

$$\widehat{u} = \mathrm{sgn}\left(\widehat{f}(\boldsymbol{t}; \widehat{\boldsymbol{\theta}}_{AIWLS})\right), \tag{28}$$

where $\mathrm{sgn}(\cdot)$ denotes the sign of a scalar. Note that, if $P_{train}(\boldsymbol{x}) = P_{test}(\boldsymbol{x})$, this classification method is equivalent to *linear discriminant analysis* (LDA) (Fisher, 1936; Duda et al., 2001), given that the class labels are $y_i \propto \{1/n_+, -1/n_-\}$, where n_+ and n_- are the numbers of positive and negative training samples, respectively. In the following, we rescale the training output values $\{y_i\}_{i=1}^n$ as such, and refer to AIWLS as *adaptive importance weighted LDA* (AIWLDA). Figure 3(b) shows an example of realizations of training and test samples, and decision boundaries obtained by AIWLDA with $\lambda = 0, 0.5, 1$. In this particular realization, $\lambda = 0.5$ or 1 seems to work better than $\lambda = 0$. However, the best value of λ depends on the realization of samples and λ needs to be optimized by a model selection method.

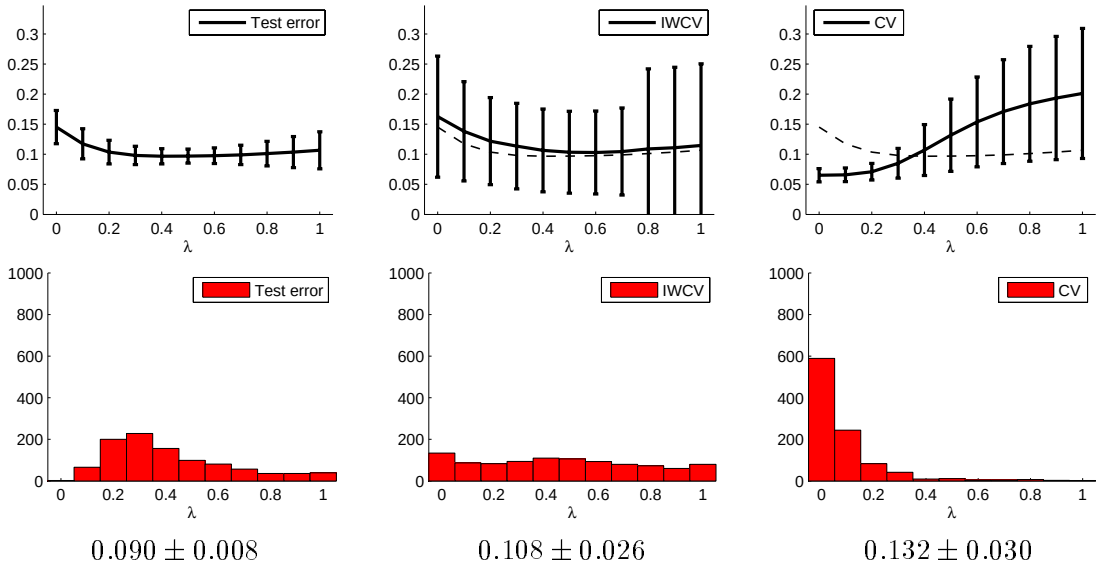


Figure 4: Top graphs: Test error (misclassification rate) and its estimates as functions of the tuning parameter λ in AIWLDA. Dashed curves in the bottom graphs depict the mean test error (i.e., the mean of the left graph) for clear comparison. Bottom graphs: Histograms of the minimizer. The numbers below the graphs are the means and standard deviations of the test error when λ is selected by each method.

Top graphs in Figure 4 depicts the mean and standard deviation of the test error (i.e., misclassification rate) and its estimate by each method over 1000 runs, as functions of the tuning parameter λ in AIWLDA. The graphs clearly show that IWCV gives much better estimates of the risk than CV.

Next we investigate the model selection performance: λ is chosen from $\{0, 0.1, 0.2, \ldots, 1\}$ so that the score of each method is minimized. Bottom graphs in Figure 4 depict the histograms of the minimizer of each score. The mean and standard deviation of the test error when λ is chosen by each method are described below the graphs. The numbers show that IWCV gives much smaller test errors than CV (the difference is significant by the *t-test* at the significance level 1%).

We also carried out the same simulation except that the training and test densities are unknown; they are estimated by maximum likelihood fitting of a single Gaussian model or a Gaussian kernel density estimator with variance determined by Silverman's rule-of-thumb bandwidth selection rule. For estimating the test input density, we draw 500 unlabeled samples following $P_{test}(\boldsymbol{x})$. The simulation results were almost identical to the known-density case with a small increase in the error (therefore we omit the detail). This implies that, for this toy classification problem, estimating the densities from data does not significantly degrade the quality of learning.

This simulation result illustrated that IWCV is useful also in covariate-shifted classification tasks.

5. Application to Brain-Computer Interface

The previous section showed that IWCV is promising in both regression and classification tasks with covariate shift. In particular, IWCV is the *only* method which can be successfully applied in covariate-shifted *classification* scenarios. In this section, we apply IWCV to the classification tasks in brain-computer interfaces (BCIs), which attracts a lot of attention these days in biomedical engineering.

A BCI is a system which allows for a direct communication from man to machine (Wolpaw et al., 2002; Dornhege et al., 2007). Cerebral electric activity is recorded via the *electroencephalogram* (EEG): electrodes attached to the scalp measure the electric signals of the brain. These signals are amplified and transmitted to the computer, which translates them into device control commands. The crucial requirement for the successful functioning of BCI is that the electric activity on the scalp surface already reflects motor intentions, i.e., the neural correlate of preparation for hand or foot movements. A BCI can detect the motor-related EEG changes and uses this information, for example, to perform a choice between two alternatives: the detection of the preparation to move the left hand leads to the choice of the first control command, whereas the right hand intention would lead to the second alternative. By this means, it is possible to operate devices which are connected to the computer.

For classification of bandpower estimates of appropriately preprocessed EEG signals (Ramoser et al., 2000; Pfurtscheller and da Silva, 1999; Lemm et al., 2005), LDA has shown to work very well (Wolpaw et al., 2002; Dornhege et al., 2004; Babiloni et al., 2000). On the other hand, strong non-stationarity effects have been often observed in brain signals between training and test sessions (Vidaurre et al., 2004; Millán, 2004; Shenoy et al., 2006), which could be regarded as an example of the covariate shift. This indicates that employing importance weighted methods could further improve the BCI recognition accuracy.

Here, we employ AIWLDA to cope with the non-stationarity (see Section 4.2 for detail). We test AIWLDA with totally 14 data sets obtained from 5 different subjects (see Table 1 for specification), where the task is binary classification of EEG signals. The experimental setting is described in more detail in the references (Blankertz et al., 2007, 2006; Sugiyama et al., 2006). Note that training samples and unlabeled/test samples are gathered in different recording sessions, so the non-stationarity in brain signals may change the distributions. On the other hand, the unlabeled samples and test samples are gathered in the same recording session; more precisely, the unlabeled samples are gathered in the first half of the session while the test samples (with labels) are collected in the latter half. Therefore, unlabeled samples may contain some information on the test input distribution, although input distributions of unlabeled and test samples are not necessarily identical since the non-stationarity in brain signals can cause a small change in distributions even within the same session. Thus, this setting realistically renders the classifier update in online BCI systems.

We estimate $p_{train}(\boldsymbol{x})$ and $p_{test}(\boldsymbol{x})$ by maximum likelihood fitting of the multi-dimensional Gaussian density with full covariance matrix. $p_{train}(\boldsymbol{x})$ is estimated using training samples and $p_{test}(\boldsymbol{x})$ is estimated using the unlabeled samples.

Table 2 describes the misclassification rates of the test samples by LDA (an existing method, which corresponds to AIWLDA with $\lambda = 0$), AIWLDA with λ chosen based on 10-fold IWCV or 10-fold CV, and AIWLDA with optimal λ. The value of λ is selected from $\{0, 0.1, 0.2, \ldots, 1.0\}$; chosen values are also described in the table. Table 2 also contains the *Kullback-Leibler (KL) divergence* (Kullback and Leibler, 1951) from the estimated training input distribution to the estimated

test input distribution. Since we want to have an accurate estimate of the KL divergence, we used the test samples for estimating the test input distribution when computing the KL divergence (cf. only unlabeled samples are used when the test input distribution is estimated for AIWLDA and IWCV). The KL values may be roughly interpreted as the *level* of the covariate shift.

First we compare OPT (AIWLDA with optimal λ) with LDA. The table shows that OPT outperforms LDA in 8 out of 14 cases, which motivates us to employ AIWLDA in BCI. Within each subject, we can observe a clear tendency that OPT outperforms LDA when the KL divergence is large, while they are comparable to each other when the KL divergence is small. This well agrees with the theory that AIWLDA can compensate for the effect of the covariate shift, while AIWLDA is reduced to plain LDA in the absence of covariate shift. Next we compare IWCV (applied to AIWLDA) with LDA. IWCV outperforms LDA for 5 cases, while the opposite case occurs only once. The table also shows that within each subject, IWCV tends to outperform LDA when the KL divergence is large. Finally, we compare IWCV with CV (applied to AIWLDA). IWCV outperforms CV for 3 cases, while the opposite case does not occur. IWCV tends to outperform CV when the KL divergence is large within each subject.

6. Discussions, Conclusions, and Future Prospects

In this paper, we discussed the model selection problem under the *covariate shift* paradigm: training input points and test input points are drawn from different distributions (i.e., $P_{train}(\boldsymbol{x}) \neq P_{test}(\boldsymbol{x})$), but the functional relation remains unchanged (i.e., $P_{train}(y|\boldsymbol{x}) = P_{test}(y|\boldsymbol{x})$). Under the covariate shift, standard model selection schemes such as cross validation (CV) are heavily biased and do not work as desired. In this paper, we therefore proposed a new variant of CV called importance weighted CV (IWCV), which is proved to be almost unbiased even under the covariate shift.

The model selection problem under the covariate shift has been studied so far. For example, a risk estimator in the context of density estimation called *Akaike's information criterion* (AIC) (Akaike, 1974) was modified to be still asymptotic unbiased (Shimodaira, 2000) and a risk estimator in linear regression called *subspace information criterion* (SIC) (Sugiyama and Ogawa, 2001) was similarly extended to be still unbiased (Sugiyama and Müller, 2005). Although these model selection criteria have rich theoretical properties, the generality of the proposed IWCV goes far beyond them: for the first time arbitrary models, arbitrary parameter learning methods, and arbitrary loss functions can be employed (see Sugiyama et al., 2006, for further discussion and simulation). Thanks to this generality, IWCV enabled us to select appropriate models even in classification tasks under the covariate shift.

We proved that IWCV is almost unbiased even under the covariate shift, which guarantees the quality of IWCV as a risk estimator. However, this does not necessarily imply the good model selection performance since the estimator has some variance. Although our experiments showed that IWCV works well in model selection under the covariate shift, it will be important to also theoretically closer investigate its model selection performance, e.g., following the lines of Stone (1977) or Altman and Léger (1997).

In theory, we assumed that the ratio of test and training input densities at training input points is known. On the other hand, they are replaced by appropriate empirical estimates in our simulations. Although the simulation results showed that the proposed method works well even when the densities are unknown, it is valuable to theoretically evaluate the effect of this replacement. Furthermore, developing a more sophisticated method of estimating the density ratio is an important issue

Table 1: Specification of BCI data.

Subject	Trial	Dim. of samples	# of training samples	# of unlabeled samples	# of test samples
1	1	3	280	112	112
1	2	3	280	120	120
1	3	3	280	35	35
2	1	3	280	113	112
2	2	3	280	112	112
2	3	3	280	35	35
3	1	3	280	91	91
3	2	3	280	112	112
3	3	3	280	30	30
4	1	6	280	112	112
4	2	6	280	126	126
4	3	6	280	35	35
5	1	2	280	112	112
5	2	2	280	112	112

Table 2: Misclassification rates for BCI data. All values are in percent. IWCV or CV refers to AIWLDA with λ chosen by 10-fold IWCV or 10-fold CV. OPT refers to AIWLDA with optimal λ. Values of chosen λ are described in the bracket (LDA is denoted as $\lambda = 0$). '$*$' in the table indicates the case where OPT is better than LDA. '$+$' is the case where IWCV outperforms LDA and '$-$' is the opposite case where LDA outperforms IWCV. '$\circ$' denotes the case where IWCV outperforms CV. KL refers to the Kullback-Leibler divergence between (estimated) training and test input distributions.

Subject	Trial	OPT	LDA	IWCV	CV	KL
1	1	* 8.7 (0.5)	9.3 (0)	$^{-}$ 10.0 (0.9)	10.0 (0.9)	0.76
1	2	* 6.2 (0.3)	8.8 (0)	8.8 (0)	8.8 (0)	1.11
1	3	4.3 (0)	4.3 (0)	4.3 (0)	4.3 (0)	0.69
2	1	40.0 (0)	40.0 (0)	$^{\circ}$ 40.0 (0)	41.3 (0.7)	0.97
2	2	* 38.7 (0.1)	39.3 (0)	$^{+}_{\circ}$ 38.7 (0.2)	39.3 (0)	1.05
2	3	25.5 (0)	25.5 (0)	25.5 (0)	25.5 (0)	0.43
3	1	* 34.4 (0.2)	36.9 (0)	$^{+}$ 34.4 (0.2)	34.4 (0.2)	2.63
3	2	* 18.0 (0.4)	21.3 (0)	$^{+}$ 19.3 (0.6)	19.3 (0.9)	2.88
3	3	* 15.0 (0.6)	22.5 (0)	$^{+}$ 17.5 (0.3)	17.5 (0.4)	1.25
4	1	* 20.0 (0.2)	21.3 (0)	21.3 (0)	21.3 (0)	9.23
4	2	2.4 (0)	2.4 (0)	2.4 (0)	2.4 (0)	5.58
4	3	6.4 (0)	6.4 (0)	6.4 (0)	6.4 (0)	1.83
5	1	21.3 (0)	21.3 (0)	21.3 (0)	21.3 (0)	0.79
5	2	* 13.3 (0.5)	15.3 (0)	$^{+}_{\circ}$ 14.0 (0.1)	15.3 (0)	2.01

to be explored (see e.g., Huang et al., 2007). Feature selection and dimensionality reduction are also important ingredients for better performance. Taking into account the covariate shift in feature selection and dimensionality reduction would be an interesting issue to be pursued, e.g., following the lines of He and Niyogi (2004) or Sugiyama (2006b).

While we focused on AIWERM and investigated the model selection performance, developing more sophisticated parameter learning methods would be an important future direction. Graph regularization techniques (Bousquet et al., 2004; Belkin and Niyogi, 2004; Hein, 2006; Chapelle et al., 2006), support vector machines (SVMs) (Vapnik, 1998; Schölkopf and Smola, 2002; Huang et al., 2007), boosting (Schapire, 2003; Meir and Rätsch, 2003) could be useful bases for further development. We note that the proposed IWCV is applicable to any parameter learning methods; even non-parametric learning methods can be employed. Therefore, IWCV may be used for model selection of newly developed learning methods in the future.

We showed experimentally that the IWCV method contributes to improving the accuracy of brain-computer interfaces (BCIs). Future studies along this line will focus on the development of a real time version of the current idea, ultimately striving for fully adaptive learning systems that can appropriately deal with various kinds of non-stationarity. In addition to BCI, there are a number of possible applications, e.g., robot control (Shelton, 2001), spam filtering (Bickel and Scheffer, 2007), and bioinformatics (Baldi et al., 1998). Applying IWCV in these application areas would be an interesting direction to be investigated.

Active learning (MacKay, 1992; Cohn et al., 1996; Fukumizu, 2000)—also referred to as *experimental design* in statistics (Kiefer, 1959; Fedorov, 1972; Pukelsheim, 1993)—is the problem of determining the location of training input points $\{\boldsymbol{x}_i\}_{i=1}^n$ so that the risk is minimized. The covariate shift naturally occurs in the active learning scenario since the training input points are generated following a user-defined distribution. For linear regression, IWLS-based active learning methods that focus on minimizing the variance of the estimator have been developed (Wiens, 2000; Sugiyama, 2006a). For general situations including classification with logistic regression models, more elaborated active learning methods which use IWERM have been developed (Kanamori and Shimodaira, 2003; Bach, 2007). In the active learning scenarios, the model has to be fixed, e.g., in the above papers, λ in AIWERM (9) is fixed to one. On the other hand, in model selection scenarios, the training input points have to be fixed and corresponding training output values have to be gathered. Thus there exists a dilemma between active learning and model selection (Sugiyama and Ogawa, 2003). An interesting future direction would be to develop a method of performing active learning and model selection at the same time, e.g., following the line of Sugiyama and Rubens (2007).

A general situation where the joint training distribution and the joint test distribution are different (i.e., $P_{train}(\boldsymbol{x}, y) \neq P_{test}(\boldsymbol{x}, y)$) is called the *sample selection bias* (see Heckman, 1979). Bayesian generative approaches to coping with such situations have been proposed when unlabeled test input points are available (Storkey and Sugiyama, 2007) or when both test input points and test output values are available (Daumé III and Marcu, 2006). However, due to the Bayesian nature, these approaches implicitly assume that the model used for learning is correctly specified. When this is not true, we may need to reasonably restrict the type of distribution change for meaningful estimations (see, e.g., Zadrozny, 2004; Fan et al., 2005; Ben-David et al., 2007; Yamazaki et al., 2007, for theoretical analyses). The covariate shift setting which we discussed in this paper could be regarded as one of such restrictions.

Another interesting restriction on the distribution change is the *class prior probability change* in classification scenarios, where the class prior probabilities are different (i.e., $P_{train}(y) \neq P_{test}(y)$),

but the class conditional distribution remains unchanged (i.e., $P_{train}(\boldsymbol{x}|y) = P_{test}(\boldsymbol{x}|y)$). Note that in this case, the functional relation generally changes (i.e., $P_{train}(y|\boldsymbol{x}) \neq P_{test}(y|\boldsymbol{x})$). SVM (Vapnik, 1998; Schölkopf and Smola, 2002) is a popular classification technique and is shown to converge to the Bayes optimal classifier as the number of training samples tends to infinity (Lin, 2002). However, this nice theoretical property is lost under the class prior probability change. To cope with this problem, SVMs are modified so that the convergence to the Bayes optimal classifier is still guaranteed under the class prior probability change (Lin et al., 2002). Our proposed IWCV idea can be similarly applied in the scenarios of class prior probability change; more specifically, if we replace the importance $p_{test}(\boldsymbol{x}_i)/p_{train}(\boldsymbol{x}_i)$ by $P_{test}(y_i)/P_{train}(y_i)$, IWCV is still almost unbiased even under the class prior probability change. Therefore, IWCV may be used for tuning the model parameters of SVMs even under the class prior probability change. Note that the setting of class prior probability change may be regarded as an extension of *imbalanced classification*, where the ratio of training samples in each class is not even (see e.g., Japkowicz, 2000; Chawla et al., 2003).

Beyond the covariate shift, learning under changing distribution has been gathering significant attention recently (e.g., Bickel, 2006; Candela et al., 2006); note also the large body of work exists in online learning, where the distribution is subject to continuous change (e.g., Robbins and Munro, 1951; Saad, 1998; LeCun et al., 1998; Murata et al., 2002). For further developing learning methods under the changing environment, it is essential to establish and share standard benchmark data sets, e.g., the projects supported by PASCAL (Candela et al., 2005) or EPSRC (Kuncheva, 2006), Common benchmark data sets can be used to evaluate the experimental performance of proposed and related methods.

Finally, the importance-weighting idea which was originally used in importance sampling (e.g., Fishman, 1996) could be applied to various statistical procedures, including resampling techniques such as *bootstrap* (Efron, 1979; Efron and Tibshirani, 1993). An interesting future direction is therefore to develop a family of importance-weighted algorithms following the spirit of this paper and to investigate their statistical properties.

Acknowledgments

The authors would like to thank Mitsuo Kawato, Kazuyuki Aihara, Olivier Chapelle, Bernhard Schölkopf, Benjamin Blankertz, Guido Dornhege, Tobias Scheffer, and Masato Okada for their fruitful comments. This work was supported in part by grants of MEXT (Grant-in-Aid for Young Scientists 17700142 and Grant-in-Aid for Scientific Research (B) 18300057) and BMBF (FKZ 01IBE01A/B). A part of this work has been done when MS was staying at University of Edinburgh, which is supported by EU Erasmus Mundus Scholarship. He would like to thank Sethu Vijayakumar for the warm hospitality.

References

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

N. Altman and C. Léger. On the optimality of prediction-based selection criteria and the convergence rates of estimators. *Journal of the Royal Statistical Society, Series B*, 59(1):205–216, 1997.

F. Babiloni, F. Cincotti, L. Lazzarini, J. d. R. Millán, J. Mourinõ, M. Varsta, J. Heikkonen, L. Bianchi, and M. G. Marciani. Linear classification of low-resolution EEG patterns produced by imagined hand movements. *IEEE Transactions on Rehabilitation Engineering*, 8(2):186–188, June 2000.

F. Bach. Active learning for misspecified generalized linear models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

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

M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56(1–3):209–239, 2004.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

S. Bickel. ECML2006 workshop on discovery challenge, 2006. URL `http://www.ecmlpkdd2006.org/challenge.html`.

S. Bickel and T. Scheffer. Dirichlet-enhanced spam filtering based on biased samples. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

B. Blankertz, G. Dornhege, M. Krauledat, and K.-R. Müller. The Berlin brain-computer interface: EEG-based communication without subject training. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):147–152, 2006.

B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio. The non-invasive Berlin brain-computer interface: Fast acquisition of effective performance in untrained subjects. *NeuroImage*, 2007. to appear.

K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

O. Bousquet, O. Chapelle, and M. Hein. Measure based regularization. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

J. Q. Candela, N. Lawrence, and A. Schwaighofer. Learning when test and training inputs have different distributions challenge, 2005. URL `http://www.pascal-network.org/Challenges/LETTIDD/`.

J. Q. Candela, N. Lawrence, A. Schwaighofer, and M. Sugiyama. NIPS2006 workshop on learning when test and training inputs have different distributions, 2006. URL `http://ida.first.fraunhofer.de/projects/different06/`.

O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, 2006.

N. Chawla, N. Japkowicz, and A. Kolcz. ICML2003 workshop on learning from imbalanced data sets, 2003. URL `http://www.site.uottawa.ca/~nat/Workshop2003/workshop2003.html`.

D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

H. Daumé III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006.

G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller. Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms. *IEEE Transactions on Biomedial Engineering*, 51(6):993–1002, June 2004.

G. Dornhege, J. d. R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, editors. *Towards Brain Computer Interfacing*. MIT Press, 2007. in preparation.

R. O. Duda, P. E. Hart, and D. G. Stor. *Pattern Classification*. Wiley, New York, 2001.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

W. Fan, I. Davidson, B. Zadrozny, and P. S. Yu. An improved categorization of classifier's sensitivity on sample selection bias. In *In Proceedings of the Fifth IEEE International Conference on Data Mining*, 2005.

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

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936.

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

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

W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Springer, Berlin, 2004.

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

X. He and P. Niyogi. Locality preserving projections. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

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

M. Hein. Uniform convergence of adaptive graph-based regularization. In *Proceedings of the 19th Annual Conference on Learning Theory*, pages 50–64, 2006.

J. Huang, A. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

N. Japkowicz. AAAI2000 workshop on learning from imbalanced data sets, 2000. URL `http://www.site.uottawa.ca/~nat/Workshop2000/workshop2000.html`.

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.

J. Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society, Series B*, 21: 272–304, 1959.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

L. Kuncheva. Classifiers ensembles for changing environments, 2006. URL `http://gow.epsrc.ac.uk/ViewGrant.aspx?GrantRef=EP/D04040X/1`.

Y. LeCun, L. Bottou, G. B. Orr, and K.-R Müller. Efficient backprop. In G. Orr and K.-R. Müller, editors, *Neural Networks: Tricks of the Trade*, number 1524 in Lecture Notes in Computer Science, pages 299–314. Springer, Berlin, 1998.

S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller. Spatio-spectral filters for improved classification of single trial EEG. *IEEE Transactions on Biomedial Engineering*, 52(9):1541–1548, Sept. 2005.

Y. Lin. Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery*, 6(3):259–275, 2002.

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

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

D. J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

R. Meir and G. Rätsch. An introduction to boosting and leveraging. In S. Mendelson and A. J. Smola, editors, *Advanced Lectures on Machine Learning*, pages 119–184. Springer, Berlin, 2003.

J. d. R. Millán. On the need for on-line learning in brain-computer interfaces. In *Proceedings of the International Joint Conference on Neural Networks*, volume 4, pages 2877–2882, Budapest, Hungary, July 2004.

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.

G. Pfurtscheller and F. H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110(11):1842–1857, Nov 1999.

F. Pukelsheim. *Optimal Design of Experiments*. Wiley, 1993.

H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–446, 2000.

H. Robbins and S. Munro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

D. Saad, editor. *On-Line Learning in Neural Networks*. Cambridge University Press, Cambridge, 1998.

R. E. Schapire. The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, and B. Yu, editors, *Nonlinear Estimation and Classification*, pages 149–172. Springer, New York, 2003.

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

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

P. Shenoy, M. Krauledat, B. Blankertz, R. P. N. Rao, and K.-R. Müller. Towards adaptive classification for BCI. *Journal of Neural Engineering*, 3:R13–R23, 2006.

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.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36:111–147, 1974.

M. Stone. Asymptotics for and against cross-validation. *Biometrika*, 64(1):29–35, 1977.

A. Storkey and M. Sugiyama. Mixture regression for covariate shift. In B. Schölkopf, J. C. Platt, and T. Hoffmann, editors, *Advances in Neural Information Processing Systems 19*, Cambridge, MA, 2007. MIT Press.

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

M. Sugiyama. Local Fisher discriminant analysis for supervised dimensionality reduction. In W. Cohen and A. Moore, editors, *Proceedings of 23rd International Conference on Machine Learning*, pages 905–912, Pittsburgh, Pennsylvannia, USA, Jun. 25–29 2006b.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. Technical Report TR06-0007, Department of Computer Science, Tokyo Institute of Technology, Sep. 2006. URL `http://www.cs.titech.ac.jp/`.

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

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

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.

M. Sugiyama and N. Rubens. A batch ensemble approach to active learning with model selection, 2007. submitted.

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

C. Vidaurre, A. Schlögl, R. Cabeza, and G. Pfurtscheller. About adaptive classifiers for brain computer interfaces. *Biomedizinische Technik*, 49(1):85–86, 2004.

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

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

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.

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.

K. Yamazaki, M. Kawanabe, S. Watanabe, M. Sugiyama, and K.-R Müller. Asymptotic Bayesian generalization error when training and test distributions are different, 2007. submitted.

B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-First International Conference on Machine Learning*, New York, NY, 2004. ACM Press.