Importance-Weighted Cross-Validation for Covariate Shift

Masashi Sugiyama[1], Benjamin Blankertz[2], Matthias Krauledat[2,3], Guido Dornhege[2], and Klaus-Robert Müller[2,3]

[1] Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan
[2] Fraunhofer FIRST.IDA, Berlin, Germany
[3] Department of Computer Science, University of Potsdam, Potsdam, Germany
sugi@cs.titech.ac.jp, {benjamin.blankertz, matthias.krauledat, guido.dornhege, klaus}@first.fhg.de,

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 used for testing. However, this assumption is not satisfied, for example, when the outside of training region is extrapolated. The situation where the training input points and test input points follow *different* distributions is called the covariate shift. Under the covariate shift, standard machine learning techniques such as empirical risk minimization or cross-validation do not work well since their unbiasedness is no longer maintained. In this paper, we propose a new method called importance-weighted cross-validation, which is still unbiased even under the covariate shift. The usefulness of our proposed method is successfully tested on toy data and furthermore demonstrated in the brain-computer interface, where strong non-stationarity effects can be seen between calibration and feedback 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., [9, 3, 5]). However, this common assumption is not fulfilled, for example, when the outside of training region is extrapolated and 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 is called the *covariate shift* [6]. For data from many applications such as off-policy reinforcement learning, bioinformatics, or brain-computer interfacing, the covariate shift phenomenon is conceivable.

In an idealized situation where the model used for learning is *correctly specified* (i.e., the learning target is included in the model), empirical risk minimization (ERM, cf. Eq.(4)) which is a typical parameter learning method still gives an

asymptotically unbiased estimator of the true parameter even under the covariate shift. However, in practical situations where the model is *misspecified* (i.e., the learning target is not included in the model), the asymptotic unbiasedness[4] does not hold anymore; ERM yields a biased estimator even asymptotically.

To illustrate this phenomenon, let us employ a toy regression problem of fitting a linear function to the sinc function (see Figure 1). Here, we consider an extrapolation problem: training input points are distributed in the left-hand side of the input domain, while test input points are distributed in the right-hand side. The density functions of the training and test input points are depicted by the solid and dashed lines in Figure 1-(A). If ordinary least-squares (OLS) (which is an ERM method with squared-loss) is used for fitting the straight line, we have a good approximation of the left-hand side of the sinc function (see Figure 1-(B)). However, this is not an appropriate function for estimating the test output values ('$\times$' in the figure). Thus, OLS results in a large test error.

Under the covariate shift with misspecified models, *importance-weighted ERM* (IWERM, cf. Eq.(6)) is shown to give an asymptotically unbiased estimator [6]. The key idea of IWERM is to weight the empirical risk according to the *importance*, which is the ratio of densities of the training and test input points. By this density ratio, the training input distribution is systematically adjusted to the test input distribution.

Figure 1-(D) depicts the learned function obtained by importance-weighted least-squares (IWLS). Compared with OLS, IWLS gives a better function for estimating the test output values; the learned function converges to the optimal function as the number of training samples goes to infinity.

The asymptotic unbiasedness can be achieved by IWERM, which may result in good estimation of the test output values, as illustrated above. However, IWERM generally yields an estimator with larger variance than ordinary ERM. This may be intuitively confirmed by the fact that OLS is the best linear unbiased estimator, i.e., having the smallest variance among all linear unbiased estimators. Therefore, IWERM may not be optimal; a slightly biased variant of IWERM with smaller variance could be better. The bias-variance trade-off may be controlled by slightly 'weakening' the importance in IWERM [6] or by adding a regularization term to IWERM. We refer to such a variance-reduced variant as *adaptive IWERM* (AIWERM, cf. Eq.(8)). AIWERM includes a tuning parameter λ ($0 \leq \lambda \leq 1$); $\lambda = 0$ corresponds to ordinary ERM (uniform weight) and $\lambda = 1$ corresponds to IWERM (weight equal to the importance).

Figure 1-(C) depicts a learned function obtained by AIWLS with $\lambda = 0.5$, which yields much better estimation of the test output values than IWLS (AIWLS with $\lambda = 1$) or OLS (AIWLS with $\lambda = 0$).

As the above simple regression example demonstrates, AIWERM can work very well given λ is chosen appropriately. However, $\lambda = 0.5$ is not always the

[4] Usually an estimator is said to be unbiased if the expectation of the estimator agrees with the true parameter. For a misspecified model, we say that an estimator is unbiased if the expectation of the estimator agrees with the optimal parameter in the model (i.e., the optimal approximation of the learning target).

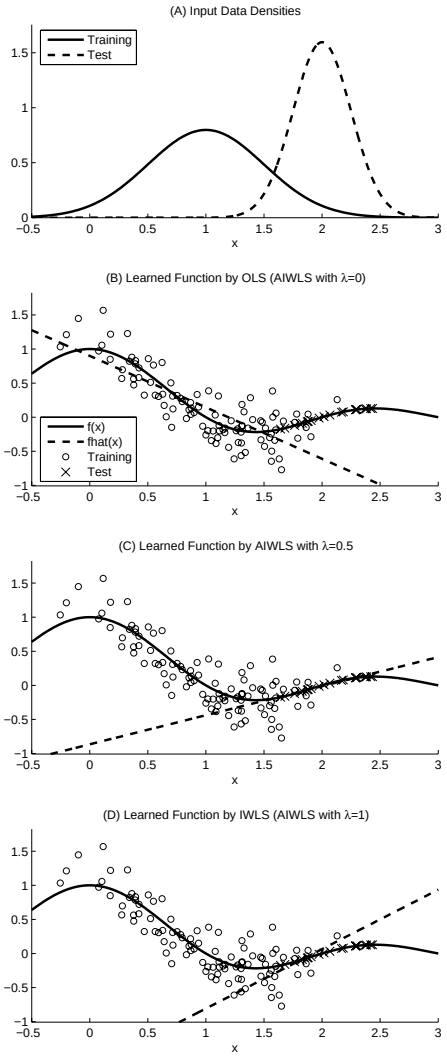

Fig. 1. An illustrative example of extrapolation by fitting a linear function. (A) The probability density functions of the training and test input points. (B)–(D) The learning target function $f(x)$ (the solid line), the noisy training samples ('o'), a learned function $\widehat{f}(x)$ (the dashed line), and the (noiseless) test samples ('×').

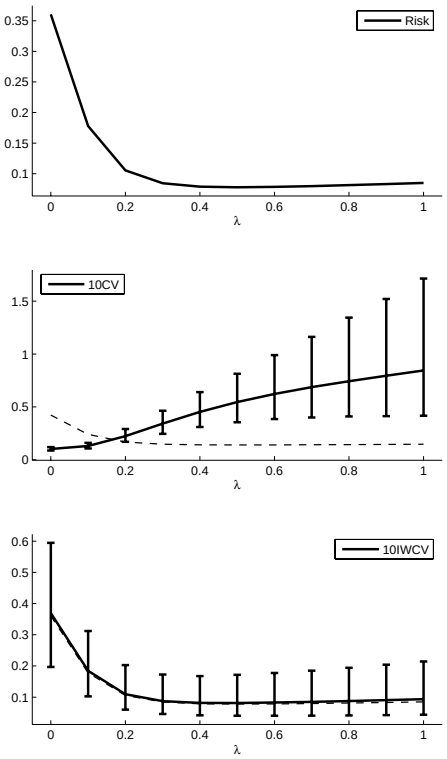

Fig. 2. True risk and its estimations as functions of the tuning parameter λ in AIWLS. Dotted curves in the bottom two graphs depict the true risk for clear comparison.

best choice; a good value of λ may depend on the learning target, used model, noise in the training samples, etc. Therefore, for enhancing generalization capability under the covariate shift, *model selection* should be carried out: set the value of the tuning parameter λ so that the estimated risk (or the estimated generalization error) is minimized.

One of the popular techniques for estimating the risk in the machine learning community is *cross-validation* (CV). CV has been shown to give an almost unbiased estimate of the risk with finite samples [5]. However, this almost unbiasedness is no longer true under the covariate shift. This phenomenon is illustrated in Figure 2, which depicts the values of the true risk and its estimates as functions of the tuning parameter λ in AIWLS (the same toy regression example of Figure 1 is still used). The dotted curves in the bottom two graphs depict the true risk for clear comparison. In this example, the true risk hits the bottom at around $\lambda = 0.5$ (see the top graph of Figure 2). On the other hand, CV gives a totally different, monotone increasing curve (see the second graph of Figure 2). As a result, CV chooses $\lambda = 0$ as the best value, which appears to be a poor choice.

To cope with this problem, we propose using a novel variant of CV called *importance-weighted* CV (IWCV). We prove that IWCV is guaranteed to give an almost unbiased estimate of the risk even under the covariate shift. The bottom graph of Figure 2 shows the estimated risk obtained by IWCV. It gives much better estimation than ordinary CV, and therefore an appropriate value of λ may be chosen by IWCV.

2 Problem Formulation

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

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 with density $p(\boldsymbol{x})$ and $y_i \in \mathcal{Y} \subset \mathbb{R}$ is a training output value following a conditional probability distribution with conditional density $r(y_i|\boldsymbol{x}_i)$.

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 $\widehat{y}$. In regression scenarios where $\mathcal{Y}$ is continuous, the squared-loss is often used.

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

On the other hand, in classification scenarios where $\mathcal{Y}$ is discrete (i.e., categorical), the following 0/1-loss is a typical choice since it corresponds to the misclassification rate.

$$\ell(\boldsymbol{x}, y, \widehat{y}) = \begin{cases} 0 \text{ if } \widehat{y} = y, \\ 1 \text{ otherwise.} \end{cases} \tag{2}$$

Although the above loss functions are independent of $\boldsymbol{x}$, the loss can generally depend on $\boldsymbol{x}$ [5].

Let us use a parameterized function $\widehat{f}(\boldsymbol{x};\boldsymbol{\theta})$ for estimating the output value y, where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^b$. The goal of supervised learning is to determine the value of the parameter $\boldsymbol{\theta}$ so that the expected loss for the test samples (i.e., the risk or the generalization error) is minimized. Let $(\boldsymbol{t}, u)$ be a test sample, where $\boldsymbol{t} \in \mathcal{X}$ is a test input point and $u \in \mathcal{Y}$ is a test output value following the conditional distribution with conditional density $r(u|\boldsymbol{t})$. Note that the conditional density $r(\cdot|\cdot)$ is the same conditional density as the training output values $\{y_i\}_{i=1}^n$. Then the risk is expressed as

$$R^{(n)} = \mathbb{E}_{\{\boldsymbol{x}_i, y_i\}_{i=1}^n, \boldsymbol{t}, u}\left[\ell\left(\boldsymbol{t}, u, \widehat{f}(\boldsymbol{t};\widehat{\boldsymbol{\theta}})\right)\right], \tag{3}$$

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 standard supervised learning theories (e.g., [9, 3, 5]), the test input point $\boldsymbol{t}$ is assumed to follow $p(\boldsymbol{x})$, which is the *same* probability density as the training input points $\{\boldsymbol{x}_i\}_{i=1}^n$. On the other hand, in this paper, we consider the situation under the *covariate shift*, i.e., the test input point $\boldsymbol{t}$ follows a probability distribution with density $q(\boldsymbol{t})$, which is *different* from $p(\boldsymbol{x})$.

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):

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

If $p(\boldsymbol{x}) = q(\boldsymbol{x})$, $\widehat{\boldsymbol{\theta}}_{ERM}$ is an asymptotically unbiased estimator of the optimal parameter. However, under the covariate shift where $p(\boldsymbol{x}) \neq q(\boldsymbol{x})$, ERM does not provide an asymptotically unbiased estimator anymore; $\widehat{\boldsymbol{\theta}}_{ERM}$ is biased even asymptotically:

$$\lim_{n\to\infty}\left\{\mathbb{E}_{\{\boldsymbol{x}_i, y_i\}_{i=1}^n}\left[\widehat{\boldsymbol{\theta}}_{ERM}\right]\right\} \neq \boldsymbol{\theta}^*, \tag{5}$$

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

Under the covariate shift, the following importance-weighted ERM (IWERM) gives an asymptotically unbiased estimator [6]:

$$\widehat{\boldsymbol{\theta}}_{IWERM} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{q(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}\ell\left(\boldsymbol{x}_i, y_i, \widehat{f}(\boldsymbol{x}_i;\boldsymbol{\theta})\right)\right], \tag{6}$$

which satisfies

$$\lim_{n\to\infty}\left\{\mathbb{E}_{\{\boldsymbol{x}_i, y_i\}_{i=1}^n}\left[\widehat{\boldsymbol{\theta}}_{IWERM}\right]\right\} = \boldsymbol{\theta}^*. \tag{7}$$

From here on, we assume that $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are known and strictly positive (i.e., non-zero) for all $\boldsymbol{x} \in \mathcal{X}$.

Although the asymptotic unbiasedness is guaranteed in IWERM, it generally has larger variance than ordinary ERM [6]. Therefore, IWERM may not be optimal; a slightly biased variant of IWERM could have much smaller variance, and thus is more accurate than plain IWERM. The bias-variance trade-off may be controlled, for example, by weakening the weight (Adaptive IWERM; AIWERM):

$$\widehat{\boldsymbol{\theta}}_{AIWERM} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{q(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)} \right)^{\lambda} \ell \left(\boldsymbol{x}_i, y_i, \widehat{f}(\boldsymbol{x}_i; \boldsymbol{\theta}) \right) \right], \tag{8}$$

where $0 \leq \lambda \leq 1$.

The above AIWERM is just examples; there may be many other possibilities for controlling the bias-variance trade-off. However, we note that the methodology we propose in this paper is valid for *any* parameter learning method.

2.3 Cross-Validation Estimate of Risk

Now we want to determine the value of the tuning parameter, say λ, so that the risk $R^{(n)}$ is minimized—but $R^{(n)}$ is inaccessible. A standard approach to coping with this problem is to prepare some candidates $\{\lambda_i\}$ of the tuning parameter, to estimate the risk for each candidate, and to choose the one with minimum estimated risk.

Cross-validation (CV) is a popular method to estimate the risk $R^{(n)}$. Let us 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 (kCV) estimate of the risk $R^{(n)}$ is given by

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

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

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

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

It is known that, if $p(\boldsymbol{x}) = q(\boldsymbol{x})$, LOOCV gives an almost unbiased estimate of the risk; more precisely, LOOCV gives an unbiased estimate of the risk with $n - 1$ samples [5].

$$\mathbb{E}_{\{\boldsymbol{x}_i, y_i\}_{i=1}^{n}} \left[\widehat{R}^{(n)}_{LOOCV} \right] = R^{(n-1)} \approx R^{(n)}. \tag{11}$$

However, this is no longer true under the covariate shift with $p(\boldsymbol{x}) \neq q(\boldsymbol{x})$. 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

Under the covariate shift, we propose using the following importance-weighted cross-validation (IWCV):

$$\widehat{R}^{(n)}_{kIWCV} = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{|\mathcal{T}_j|}\sum_{(\boldsymbol{x},y)\in\mathcal{T}_j}\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\ell\left(\boldsymbol{x},y,\widehat{f}_{\mathcal{T}_j}(\boldsymbol{x})\right), \tag{12}$$

or

$$\widehat{R}^{(n)}_{LOOIWCV} = \frac{1}{n}\sum_{j=1}^{n}\frac{q(\boldsymbol{x}_j)}{p(\boldsymbol{x}_j)}\ell\left(\boldsymbol{x}_j,y_j,\widehat{f}_j(\boldsymbol{x}_j)\right). \tag{13}$$

Below, we prove that LOOIWCV gives an almost unbiased estimate of the risk even under the covariate shift (its proof is given in a separate technical report [8]).

Lemma 1

$$\mathbb{E}_{\{\boldsymbol{x}_i,y_i\}_{i=1}^{n}}\left[\widehat{R}^{(n)}_{LOOIWCV}\right] = R^{(n-1)}. \tag{14}$$

This lemma shows that the simple variant of CV called IWCV 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 is larger than LOOIWCV.

The density ratio $q(\boldsymbol{x})/p(\boldsymbol{x})$ also appears in *importance sampling*; an expectation $\mathbb{E}_{\boldsymbol{t}}[f(\boldsymbol{t})]$ with $\boldsymbol{t}\sim q(\boldsymbol{x})$ is computed by an equivalent quantity $\mathbb{E}_{x}[f(\boldsymbol{x})q(\boldsymbol{x})/p(\boldsymbol{x})]$ with $\boldsymbol{x}\sim p(\boldsymbol{x})$, where $p(\boldsymbol{x})$ is chosen so that the variance is minimized. Therefore, the proposed IWCV method could be regarded as an application of the importance sampling identity in the CV framework. We expect that the relation between importance sampling and covariate shift may be further discussed in the context of *active learning* [7], where the training input density $p(\boldsymbol{x})$ is designed by users so that the risk is minimized.

A weighted CV scheme has also been studied in robust statistics [1], where the effect of outliers in the CV score is deemphasized by assigning smaller weight to outliers. In the proposed IWCV scheme, the CV score is weighted by the density ratio, by which the difference between $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ can be systematically adjusted. Therefore, although using a weighted scheme in CV is a common feature, the aim is essentially different; we may even combine two schemes.

4 A Numerical Example

In this section, we experimentally investigate how IWCV works using a simple one-dimensional regression dataset (see Figure 1). Let the training and test input densities be $p(x) = \phi_{1,(1/2)^2}(x)$ and $q(x) = \phi_{2,(1/4)^2}(x)$, where $\phi_{\mu,c^2}(x)$ denotes the normal density with mean μ and variance c^2. This setting implies that we are considering an extrapolation problem (see Figure 1-(A)). We create the output

Table 1. Extrapolation in the toy dataset. The mean and standard deviation of the test error obtained by each method are described. For reference, the test error obtained with the optimal λ (i.e., the minimum test error) is described as 'OPT'.

10CV	10IWCV	OPT
0.360 ± 0.108	0.086 ± 0.041	0.073 ± 0.023

value y_i following $\phi_{f(x)1,(1/4)^2}(x)$, where $f(x) = \mathrm{sinc}(x)$. We use a simple linear model for learning:

$$\widehat{f}(x;\theta_0,\theta_1) = \theta_0 + \theta_1 x, \tag{15}$$

where the parameters are learned by adaptive importance-weighted least-squares (AIWLS):

$$\underset{\theta_0,\theta_1}{\mathrm{argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{q(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)} \right)^{\lambda} \left(\widehat{f}(\boldsymbol{x}_i;\theta_0,\theta_1) - y_i \right)^2 \right]. \tag{16}$$

Figure 1 (B)–(D) show the true function, a realization of training samples, learned functions by AIWLS with $\lambda = 0, 0.5, 1$, and a realization of (noiseless) test samples. For this particular case, $\lambda = 0.5$ seems to work well.

Figure 2 depicts the means and standard deviations of the true risk and its estimates by 10-fold CV and 10-fold IWCV over 1000 runs, as functions of the tuning parameter λ in AIWLS. The graphs show that IWCV gives much accurate estimates of the risk than ordinary CV; the unbiasedness of IWCV is well satisfied and the variance of IWCV seems to be reasonable.

We then choose λ from $\{0, 0.1, 0.2, \ldots, 1\}$ so that the ordinary CV score or the IWCV score is minimized. The means and standard deviations of the test error finally obtained by ordinary CV and IWCV over 1000 runs are described in Table 1. The table shows that IWCV gives much smaller test errors than ordinary CV; the p-value between ordinary CV and IWCV by the t-test is far less than 0.01, stating that IWCV significantly outperforms ordinary CV. 'OPT' in the table shows the test error when λ is chosen optimally, i.e., so that the true test error is minimized. The result shows that the performance of IWCV is rather close to the optimal choice.

5 Application to Brain-Computer Interface

In this section, we apply IWCV to brain-computer interface (BCI) data.

BCI is a system which allows for a direct dialog between man and machine [11]. 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, 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 appropriately preprocessed EEG signals linear discriminant analysis (LDA) [3] has shown to work very well [2]. On the other hand, strong non-stationarity effects have been observed in brain signals between calibration and feedback sessions [10], which could be regarded as an example of the covariate shift. Therefore, it is expected that some importance-weighted method could further improve the BCI recognition accuracy.

LDA is actually equivalent to least-square fitting of a linear model using binary labels $y_i = \pm 1$ [3]. Here we use its variant called adaptive importance-weighted LDA (AIWLDA):

$$\underset{\theta_0,\boldsymbol{\theta}}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{q(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)} \right)^{\lambda} \left(\theta_0 + \boldsymbol{\theta}^\top \boldsymbol{x}_i - y_i \right)^2 \right]. \tag{17}$$

We test the above method with totally 14 data sets obtained from 5 different subjects (see Table 2). In BCI, the densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are unknown. Here we estimate them by fitting the mixture of 5 Gaussians by the EM algorithm. $p(\boldsymbol{x})$ is estimated using training samples and $q(\boldsymbol{x})$ is estimated using unlabeled samples from the feedback period. The unlabeled samples are taken from the first half of each feedback period, herewith rendering the conditions for a BCI application realistic. This corresponds to an update of the used classifier in the second half of the experiment.

The misclassification rates for test samples by LDA (existing method which corresponds to AIWLDA with $\lambda = 0$) and AIWLDA with λ chosen by 10IWCV are given in Table 2. The results show that for the subjects 1 and 3, the combination of AIWLDA and 10IWCV highly improves the recognition accuracy over plain LDA. The accuracy is unchanged for the subjects 2 and 4, and comparable for the subject 5. Overall, the proposed method outperforms LDA for 5 out of 14 data sets and being outperformed for 1 data set.

Note that the degree of non-stationarity is highly subject specific. While—as expected—our method for compensating covariate shift effects yields highly significant improvements for some subjects, others exhibit no change due to the rather stationary nature of their brain signals.

6 Conclusions

In this paper, we discussed the supervised learning problem under the covariate shift paradigm: training input points and test input points are drawn from different distributions. Future studies will focus on the development of a realtime version of the current idea in order to ultimately obtain a fully adaptive learning system.

We acknowledge partial financial supports from MEXT (Grant-in-Aid for Young Scientists 17700142) and BMBF (FKZ 01IBE01A/B).

Table 2. Misclassification rates for brain computer interface. All values are in percent. The values of the better method are described using a bold face.

Subject	Trial	# of training samples	# of unlabeled samples	# of test samples	LDA	AIWLDA + 10IWCV	AIWLDA + OPT
1	1	280	112	112	9.8	**8.0**	8.0
1	2	280	120	120	10.8	10.8	6.7
1	3	280	35	35	5.7	**2.9**	2.9
2	1	280	113	112	43.4	43.4	43.4
2	2	280	112	112	38.5	38.5	38.5
2	3	280	35	35	28.6	28.6	28.6
3	1	280	91	91	39.6	**38.5**	37.4
3	2	280	112	112	22.3	**19.6**	19.6
3	3	280	30	30	20.0	20.0	20.0
4	1	280	112	112	24.1	24.1	23.2
4	2	280	126	126	2.4	2.4	2.4
4	3	280	35	35	8.6	8.6	8.6
5	1	280	112	112	**22.3**	25.0	22.3
5	2	280	112	112	12.5	**11.6**	10.7

References

1. C. Agostinelli. Robust model selection by cross-validation via weighted likelihood methodology. Technical Report 1999.37, Dipartimento di Scienze Statistiche, Università di Padova, 1999.
2. B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio. The Berlin brain-computer interface: Report from the feedback sessions. Technical Report 1, Fraunhofer FIRST, 2005.
3. R. O. Duda, P. E. Hart, and D. G. Stor. *Pattern Classification*. Wiley, New York, 2001.
4. J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–162, 1979.
5. B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
6. 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.
7. M. Sugiyama. Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research*, 7(Jan):141–166, 2006.
8. M. Sugiyama, B. Blankertz, M. Krauledat, G. Dornhege, and K.-R. Müller. Importance-weighted cross-validation for covariate shift. Technical Report TR06-0002, Department of Computer Science, Tokyo Institute of Technology, Feb. 2006.
9. V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
10. C. Vidaurre, A. Schlögl, R. Cabeza, and G. Pfurtscheller. About adaptive classifiers for brain computer interfaces. *Biomedizinische Technik*, 49(1):85–86, 2004.
11. 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.