Density Ratio Estimation: A Comprehensive Review

Masashi Sugiyama, Tokyo Institute of Technology (sugi@cs.titech.ac.jp)
Taiji Suzuki, The University of Tokyo (s-taiji@stat.t.u-tokyo.ac.jp)
Takafumi Kanamori, Nagoya University (kanamori@is.nagoya-u.ac.jp)

Abstract

Density ratio estimation has attracted a great deal of attention in the statistics and machine learning communities since it can be used for solving various statistical data processing tasks such as non-stationarity adaptation, two-sample test, outlier detection, independence test, feature selection/extraction, independent component analysis, causal inference, and conditional probability estimation. When estimating the density ratio, it is preferable to avoid estimating densities since density estimation is known to be a hard problem. In this paper, we give a comprehensive review of density ratio estimation methods based on moment matching, probabilistic classification, and ratio matching.

1 Introduction

Recently, a new general framework of statistical data processing based on the *ratio* of probability densities has been developed (Sugiyama et al., 2009; Sugiyama et al., 2011). This density ratio framework includes various statistical data processing tasks such as non-stationarity adaptation (Shimodaira, 2000; Zadrozny, 2004; Sugiyama & Müller, 2005; Sugiyama et al., 2007; Quiñonero-Candela et al., 2009; Sugiyama et al., 2010d), outlier detection (Hido et al., 2008; Smola et al., 2009; Hido et al., 2010), change detection in time series (Kawahara & Sugiyama, 2009), conditional density estimation (Sugiyama et al., 2010c), and probabilistic classification (Sugiyama, 2010).

Furthermore, *mutual information*—which plays a central role in information theory (Cover & Thomas, 2006)—can be estimated via density ratio estimation (Suzuki et al., 2008; Suzuki et al., 2009b). Since mutual information is a measure of statistical independence between random variables, density ratio estimation can be used also for variable selection (Suzuki et al., 2009a), dimensionality reduction (Suzuki & Sugiyama, 2010), independent component analysis (Suzuki & Sugiyama, 2009), and causal inference (Yamada & Sugiyama, 2010). Thus, density ratio estimation is a promising versatile tool for statistical data processing.

A naive approach to estimating the density ratio is to separately estimate the densities corresponding to the numerator and denominator of the ratio, and then take the ratio of the estimated densities. However, this naive approach is not reliable in high-dimensional problems since division by an estimated quantity can magnify the estimation error. To overcome this drawback, various approaches to directly estimating the density ratio without going through density estimation have been explored recently, including the *moment matching approach* (Gretton et al., 2009), the *probabilistic classification approach* (Qin, 1998; Cheng & Chu, 2004; Bickel et al., 2007), and the *ratio matching approach* (Sugiyama et al., 2008; Kanamori et al., 2009a; Tsuboi et al., 2009; Yamada & Sugiyama, 2009; Yamada et al., 2010). The purpose of this paper is to provide a comprehensive review of such direct density-ratio estimation methods.

The problem of density ratio estimation addressed in this paper is formulated as follows. Let $\mathcal{X}$ ($\subset \mathbb{R}^d$) be the data domain, and suppose we are given independent and identically distributed (i.i.d.) samples $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ from a distribution with density $p_{\mathrm{nu}}^*(\boldsymbol{x})$ and i.i.d. samples $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$ from another distribution with density $p_{\mathrm{de}}^*(\boldsymbol{x})$.

$$\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{nu}}^*(\boldsymbol{x}) \quad \text{and} \quad \{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{de}}^*(\boldsymbol{x}).$$

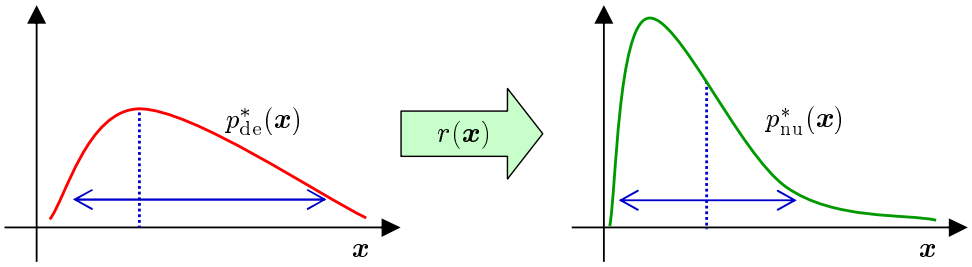

Figure 1: Matching the moments of $r(\boldsymbol{x})p^*_{\mathrm{de}}(\boldsymbol{x})$ with those of $p^*_{\mathrm{nu}}(\boldsymbol{x})$.

We assume that $p^*_{\mathrm{de}}(\boldsymbol{x})$ is strictly positive over the domain $\mathcal{X}$. The goal is to estimate the density ratio

$$r^*(\boldsymbol{x}) := \frac{p^*_{\mathrm{nu}}(\boldsymbol{x})}{p^*_{\mathrm{de}}(\boldsymbol{x})}$$

from samples $\{\boldsymbol{x}^{\mathrm{nu}}_i\}_{i=1}^{n_{\mathrm{nu}}}$ and $\{\boldsymbol{x}^{\mathrm{de}}_j\}_{j=1}^{n_{\mathrm{de}}}$. 'nu' and 'de' indicate 'numerator' and 'denominator', respectively.

2 Moment Matching Approach

In this section, we describe the *moment matching* approach to density ratio estimation.

2.1 Preliminaries

Suppose that a one-dimensional random variable x is drawn from a probability distribution with density $p^*(x)$. Then the k-th order moment of x about the origin is defined by $\int x^k p^*(x)\mathrm{d}x$. Note that two distributions are equivalent if and only if all moments (i.e., for $k = 1, 2, \ldots$) agree with each other.

The moment matching approach to density ratio estimation tries to match the moments of $p^*_{\mathrm{nu}}(\boldsymbol{x})$ and $p^*_{\mathrm{de}}(\boldsymbol{x})$ via a 'transformation' function $r(\boldsymbol{x})$. More specifically, using the true density ratio $r^*(\boldsymbol{x})$, $p^*_{\mathrm{nu}}(\boldsymbol{x})$ can be expressed as

$$p^*_{\mathrm{nu}}(\boldsymbol{x}) = r^*(\boldsymbol{x})p^*_{\mathrm{de}}(\boldsymbol{x}).$$

Thus, for a density ratio model $r(\boldsymbol{x})$, matching the moments of $p^*_{\mathrm{nu}}(\boldsymbol{x})$ and $r(\boldsymbol{x})p^*_{\mathrm{de}}(\boldsymbol{x})$ leads to the true density ratio $r^*(\boldsymbol{x})$. A schematic illustration of the moment matching approach is described in Figure 1.

2.2 Finite-Order Approach

The simplest implementation of moment matching would be to match the first-order moment (i.e., the mean):

$$\operatorname*{argmin}_r \left\| \int \boldsymbol{x} r(\boldsymbol{x}) p^*_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int \boldsymbol{x} p^*_{\mathrm{nu}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm. Its non-linear variant can be obtained using some non-linear function $\phi(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ as

$$\operatorname*{argmin}_r \left(\int \phi(\boldsymbol{x}) r(\boldsymbol{x}) p^*_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int \phi(\boldsymbol{x}) p^*_{\mathrm{nu}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right)^2.$$

This non-linear method can be easily extended to multiple components by using a vector-valued function $\boldsymbol{\phi}(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^m$ as

$$\operatorname*{argmin}_r \mathrm{MM}'(r), \quad \text{where} \quad \mathrm{MM}'(r) := \left\| \int \boldsymbol{\phi}(\boldsymbol{x}) r(\boldsymbol{x}) p^*_{\mathrm{de}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int \boldsymbol{\phi}(\boldsymbol{x}) p^*_{\mathrm{nu}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right\|^2,$$

where 'MM' stands for 'moment matching'. Let us ignore the irrelevant constant in $\mathrm{MM}'(r)$, and define the rest as $\mathrm{MM}(r)$:

$$\mathrm{MM}(r) := \left\| \int \boldsymbol{\phi}(\boldsymbol{x}) r(\boldsymbol{x}) p^*_{\mathrm{de}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right\|^2 - 2 \left\langle \int \boldsymbol{\phi}(\boldsymbol{x}) r(\boldsymbol{x}) p^*_{\mathrm{de}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \int \boldsymbol{\phi}(\boldsymbol{x}) p^*_{\mathrm{nu}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right\rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product.

In practice, the expectations over $p^*_{\mathrm{nu}}(\boldsymbol{x})$ and $p^*_{\mathrm{de}}(\boldsymbol{x})$ in $\mathrm{MM}(r)$ are replaced by sample averages. That is, for an n_{de}-dimensional vector $\boldsymbol{r}^*_{\mathrm{de}} := (r^*(\boldsymbol{x}^{\mathrm{de}}_1), \ldots, r^*(\boldsymbol{x}^{\mathrm{de}}_{n_{\mathrm{de}}}))^\top$ where $^\top$ denotes the transpose, an estimator $\widehat{\boldsymbol{r}}_{\mathrm{de}}$ of $\boldsymbol{r}^*_{\mathrm{de}}$ can be obtained by solving the following optimization problem.

$$\widehat{\boldsymbol{r}}_{\mathrm{de}} := \operatorname*{argmin}_{\boldsymbol{r} \in \mathbb{R}^{n_{\mathrm{de}}}} \widehat{\mathrm{MM}}(\boldsymbol{r}), \quad \text{where} \quad \widehat{\mathrm{MM}}(\boldsymbol{r}) := \frac{1}{n^2_{\mathrm{de}}} \boldsymbol{r}^\top \boldsymbol{\Phi}^\top_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{de}} \boldsymbol{r} - \frac{2}{n_{\mathrm{de}} n_{\mathrm{nu}}} \boldsymbol{r}^\top \boldsymbol{\Phi}^\top_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{nu}} \mathbf{1}_{n_{\mathrm{nu}}}. \quad (1)$$

$\mathbf{1}_n$ denotes the n-dimensional vector with all ones. $\boldsymbol{\Phi}_{\mathrm{nu}}$ and $\boldsymbol{\Phi}_{\mathrm{de}}$ are the $t \times n_{\mathrm{nu}}$ and $t \times n_{\mathrm{de}}$ *design matrices* defined by $\boldsymbol{\Phi}_{\mathrm{nu}} := (\boldsymbol{\phi}(\boldsymbol{x}^{\mathrm{nu}}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}^{\mathrm{de}}_{n_{\mathrm{nu}}}))$ and $\boldsymbol{\Phi}_{\mathrm{de}} := (\boldsymbol{\phi}(\boldsymbol{x}^{\mathrm{de}}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}^{\mathrm{de}}_{n_{\mathrm{de}}}))$, respectively. Taking the derivative of the objective function (1) with respect to $\boldsymbol{r}$ and setting it to zero, we have

$$\frac{2}{n^2_{\mathrm{de}}} \boldsymbol{\Phi}^\top_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{de}} \boldsymbol{r} - \frac{2}{n_{\mathrm{de}} n_{\mathrm{nu}}} \boldsymbol{\Phi}^\top_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{nu}} \mathbf{1}_{n_{\mathrm{nu}}} = \mathbf{0}_t,$$

where $\mathbf{0}_t$ denotes the t-dimensional vector with all zeros. Solving this equation with respect to $\boldsymbol{r}$, we can obtain the solution analytically as

$$\widehat{\boldsymbol{r}}_{\mathrm{de}} = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}} (\boldsymbol{\Phi}^\top_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{de}})^{-1} \boldsymbol{\Phi}^\top_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{nu}} \mathbf{1}_{n_{\mathrm{nu}}}.$$

One may add a normalization constraint $\frac{1}{n_{\mathrm{de}}} \mathbf{1}^\top_{n_{\mathrm{de}}} \boldsymbol{r} = 1$ to the optimization problem (1). Then the optimization problem becomes a linearly constrained quadratic program. Thus, a numerical solver may be needed to compute the solution. Furthermore, a non-negativity constraint $\boldsymbol{r} \geq \mathbf{0}_{n_{\mathrm{de}}}$ and/or an upper bound for a positive constant B (i.e., $\boldsymbol{r} \leq B \mathbf{1}_{n_{\mathrm{de}}}$) may also be incorporated in the optimization problem (1), where inequalities for vectors are applied in the element-wise manner. Even with these modifications, the optimization problem is still a linearly constrained quadratic program, so its solution can be numerically computed by standard optimization software.

The above moment-matching method gives an estimate of the density ratio values at the denominator sample points $\{\boldsymbol{x}^{\mathrm{de}}_j\}^{n_{\mathrm{de}}}_{j=1}$. If one wants to estimate the entire ratio function $r^*(\boldsymbol{x})$, the following linear density-ratio model may be used instead (Kanamori et al., 2009b):

$$r(\boldsymbol{x}) = \boldsymbol{\psi}(\boldsymbol{x})^\top \boldsymbol{\theta}, \quad (2)$$

where $\boldsymbol{\psi}(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^b$ is a basis function vector and $\boldsymbol{\theta}$ ($\in \mathbb{R}^b$) is a parameter vector. We assume that the basis functions are non-negative: $\boldsymbol{\psi}(\boldsymbol{x}) \geq \mathbf{0}_b$. Then model outputs at $\{\boldsymbol{x}^{\mathrm{de}}_j\}^{n_{\mathrm{de}}}_{j=1}$ are expressed in terms of the parameter vector $\boldsymbol{\theta}$ as

$$(r(\boldsymbol{x}^{\mathrm{de}}_1), \ldots, r(\boldsymbol{x}^{\mathrm{de}}_{n_{\mathrm{de}}}))^\top = \boldsymbol{\Psi}^\top_{\mathrm{de}} \boldsymbol{\theta},$$

where $\boldsymbol{\Psi}_{\mathrm{de}}$ is the $b \times n_{\mathrm{de}}$ *design matrix* defined by $\boldsymbol{\Psi}_{\mathrm{de}} := (\boldsymbol{\psi}(\boldsymbol{x}_1^{\mathrm{de}}), \ldots, \boldsymbol{\psi}(\boldsymbol{x}_{n_{\mathrm{de}}}^{\mathrm{de}}))$. Then, following Eq.(1), the parameter $\boldsymbol{\theta}$ is learned as follows.

$$\widehat{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^b} \left[\frac{1}{n_{\mathrm{de}}^2} \boldsymbol{\theta}^\top \boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{de}}^\top \boldsymbol{\Phi}_{\mathrm{de}} \boldsymbol{\Psi}_{\mathrm{de}}^\top \boldsymbol{\theta} - \frac{2}{n_{\mathrm{de}} n_{\mathrm{nu}}} \boldsymbol{\theta}^\top \boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{de}}^\top \boldsymbol{\Phi}_{\mathrm{nu}} \mathbf{1}_{n_{\mathrm{nu}}} \right]. \tag{3}$$

Taking the derivative of the above objective function with respect to $\boldsymbol{\theta}$ and setting it to zero, we have the solution $\widehat{\boldsymbol{\theta}}$ analytically as

$$\widehat{\boldsymbol{\theta}} = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}} (\boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{de}}^\top \boldsymbol{\Phi}_{\mathrm{de}} \boldsymbol{\Psi}_{\mathrm{de}}^\top)^{-1} \boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{\Phi}_{\mathrm{de}}^\top \boldsymbol{\Phi}_{\mathrm{nu}} \mathbf{1}_{n_{\mathrm{nu}}}.$$

One may include a normalization constraint, a non-negativity constraint (given that the basis function is non-negative), and a regularization constraint to the optimization problem (3):

$$\frac{1}{n_{\mathrm{de}}} \mathbf{1}_{n_{\mathrm{de}}}^\top \boldsymbol{\Psi}_{\mathrm{de}}^\top \boldsymbol{\theta} = 1, \quad \boldsymbol{\theta} \geq \mathbf{0}_b, \quad \text{and} \quad \boldsymbol{\theta} \leq B \mathbf{1}_b.$$

Then the optimization problem becomes a linearly constrained quadratic program, whose solution can be obtained by a standard numerical solver.

The upper-bound parameter B, which works as a regularizer, may be optimized by *cross-validation* (CV). That is, the numerator and denominator samples $D^{\mathrm{nu}} = \{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $D^{\mathrm{de}} = \{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$ are first divided into T disjoint subsets $\{D_t^{\mathrm{nu}}\}_{t=1}^T$ and $\{D_t^{\mathrm{de}}\}_{t=1}^T$, respectively. Then a density ratio estimator $\widehat{r}_t(\boldsymbol{x})$ is obtained from $D^{\mathrm{nu}} \backslash D_t^{\mathrm{nu}}$ and $D^{\mathrm{de}} \backslash D_t^{\mathrm{de}}$ (i.e., all samples without D_t^{nu} and D_t^{de}), and its moment matching error is computed for the hold-out samples D_t^{nu} and D_t^{de}:

$$\widetilde{\mathrm{MM}}_t(\widehat{r}) := \left(\frac{1}{|D_t^{\mathrm{de}}|} \sum_{\boldsymbol{x}^{\mathrm{de}} \in D_t^{\mathrm{de}}} \boldsymbol{\phi}(\boldsymbol{x}^{\mathrm{de}}) \widehat{r}_t(\boldsymbol{x}^{\mathrm{de}}) \right)^2 - \frac{2}{|D_t^{\mathrm{de}}||D_t^{\mathrm{nu}}|} \left(\sum_{\boldsymbol{x}^{\mathrm{de}} \in D_t^{\mathrm{de}}} \boldsymbol{\phi}(\boldsymbol{x}^{\mathrm{de}}) \widehat{r}_t(\boldsymbol{x}^{\mathrm{de}}) \right)^\top \left(\sum_{\boldsymbol{x}^{\mathrm{nu}} \in D_t^{\mathrm{nu}}} \boldsymbol{\phi}(\boldsymbol{x}^{\mathrm{nu}}) \right),$$

where $|D|$ denotes the number of elements in the set D. This procedure is repeated for $t = 1, \ldots, T$, and the average of the above hold-out moment matching error over all t is computed.

$$\widetilde{\mathrm{MM}} := \frac{1}{T} \sum_{t=1}^T \widetilde{\mathrm{MM}}_t.$$

Then the upper-bound parameter B that minimizes $\widetilde{\mathrm{MM}}$ is chosen. Availability of CV would be one of the advantages of the inductive method (i.e., learning the entire density-ratio function).

2.3 Infinite-Order Approach

Matching a finite number of moments does not necessarily result in the true density ratio function $r^*(\boldsymbol{x})$, even if infinitely many samples are available. In order to guarantee that the true density ratio function can always be obtained in the large-sample limit, all moments up to the infinite order need to be matched.

Kernel mean matching (KMM) allows one to efficiently match all the moments (Gretton et al., 2009). The basic idea of KMM is common to the above finite-order approach, but a *universal reproducing kernel* $K(\boldsymbol{x}, \boldsymbol{x}')$ (Steinwart, 2001) is used as a non-linear transformation. The *Gaussian kernel*

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2} \right) \tag{4}$$

is an example of universal reproducing kernels. It has been shown that the solution of the following optimization problem agrees with the true density ratio (Gretton et al., 2009):

$$\min_{r\in\mathcal{H}} \left\| \int K(\boldsymbol{x},\cdot)p^*_{\mathrm{nu}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int K(\boldsymbol{x},\cdot)r(\boldsymbol{x})p^*_{\mathrm{de}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right\|^2_{\mathcal{H}},$$

where $\mathcal{H}$ denotes a universal reproducing kernel Hilbert space and $\|\cdot\|_{\mathcal{H}}$ denotes its norm.

An empirical version of the above problem is reduced to

$$\min_{\boldsymbol{r}\in\mathbb{R}^{n_{\mathrm{de}}}} \left[\frac{1}{n^2_{\mathrm{de}}} \boldsymbol{r}^\top \boldsymbol{K}_{\mathrm{de,de}} \boldsymbol{r} - \frac{2}{n_{\mathrm{de}} n_{\mathrm{nu}}} \boldsymbol{r}^\top \boldsymbol{K}_{\mathrm{de,nu}} \mathbf{1}_{n_{\mathrm{nu}}} \right],$$

where $\boldsymbol{K}_{\mathrm{de,nu}}$ and $\boldsymbol{K}_{\mathrm{de,de}}$ denote the Gram matrices defined by $[\boldsymbol{K}_{\mathrm{de,nu}}]_{i,j} = K(\boldsymbol{x}^{\mathrm{de}}_i, \boldsymbol{x}^{\mathrm{nu}}_j)$ and $[\boldsymbol{K}_{\mathrm{de,de}}]_{i,i'} = K(\boldsymbol{x}^{\mathrm{de}}_i, \boldsymbol{x}^{\mathrm{de}}_{i'})$, respectively. In the same way as the finite-order case, the solution can be obtained analytically as

$$\widehat{\boldsymbol{r}}_{\mathrm{de}} = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}} \boldsymbol{K}^{-1}_{\mathrm{de,de}} \boldsymbol{K}_{\mathrm{de,nu}} \mathbf{1}_{n_{\mathrm{nu}}}. \tag{5}$$

If necessary, one may include a non-negativity constraint, a normalization constraint, and an upper bound in the same way as the finite-order case. Then the solution can be numerically obtained by solving a linearly constrained quadratic programming problem.

For the linear density-ratio model (2), an inductive variant of KMM is formulated as

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^b} \left[\frac{1}{n^2_{\mathrm{de}}} \boldsymbol{\theta}^\top \boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{K}_{\mathrm{de,de}} \boldsymbol{\Psi}^\top_{\mathrm{de}} \boldsymbol{\theta} - \frac{2}{n_{\mathrm{de}} n_{\mathrm{nu}}} \boldsymbol{\theta}^\top \boldsymbol{\Psi}_{\mathrm{de}} \boldsymbol{K}_{\mathrm{de,nu}} \mathbf{1}_{n_{\mathrm{nu}}} \right].$$

As shown above, KMM utilizes universal reproducing kernels such as the Gaussian kernel (4) to efficiently match all the moments. Theoretically, KMM is consistent for any universal reproducing kernels. However, its practical performance heavily depends on the choice of kernels such as the Gaussian width σ, and such kernel parameters cannot be simply optimized by cross-validation even in the induction cases. This is because one is not finding a Gaussian width value that matches the moments well. Thus, optimizing σ over the moment matching criterion may not be appropriate as a model selection strategy. A popular heuristic is to use the *median distance* between samples as the Gaussian width σ (Schölkopf & Smola, 2002). However, there seems no strong justification for this heuristic.

2.4 Remarks

Density ratio estimation by moment matching can successfully avoid density estimation.

The finite-order moment matching method (Section 2.2) is simple and computationally efficient, if the number of matching moments is kept reasonably small. However, the finite-order approach is not necessarily consistent. On the other hand, the infinite-order moment matching method (Section 2.3), *kernel mean matching* (KMM), can efficiently match all the moments by making use of universal reproducing kernels. Indeed, KMM has an excellent theoretical property that it is consistent. However, KMM has a limitation in model selection—there is no known method for determining the kernel parameter (such as the Gaussian kernel width). A popular heuristic of setting the Gaussian width to the median distance between samples would be useful in some cases, but this is perhaps not always reasonable.

In this section, moment matching was performed in terms of the squared norm, which led to an analytic-form solution (if no constraint is imposed). As shown in Kanamori et al. (2009b),

moment matching can be generalized to various divergences. Such a generalized KMM method actually has a close connection with the *ratio matching approach* explained in Section 4. However, the ratio matching approach is more preferable than the generalized KMM approach because of the following reasons:

- The ratio matching approach is equipped with a natural CV procedure for model selection (Sugiyama et al., 2008; Kanamori et al., 2009a). Thus no heuristic is required for choosing the Gaussian width.
- The ratio matching approach is proved to be numerically more stable than the KMM approach in terms of *condition numbers* (Kanamori et al., 2009b).

3 Probabilistic Classification Approach

In this section, we describe a framework of density ratio estimation through *probabilistic classification*.

3.1 Preliminaries

The basic idea of the probabilistic classification approach is to learn a probabilistic classifier which separates samples $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ drawn from $p_{\mathrm{nu}}^*(\boldsymbol{x})$ and samples $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$ drawn from $p_{\mathrm{de}}^*(\boldsymbol{x})$ (Qin, 1998; Cheng & Chu, 2004; Bickel et al., 2007).

Let us assign a label $y = +1$ to $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $y = -1$ to $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$, respectively. Then the two densities are written as $p_{\mathrm{nu}}^*(\boldsymbol{x}) = p^*(\boldsymbol{x}|y = +1)$ and $p_{\mathrm{de}}^*(\boldsymbol{x}) = p^*(\boldsymbol{x}|y = -1)$, respectively. Note that y is regarded as a random variable here. An application of Bayes' theorem,

$$p^*(\boldsymbol{x}|y) = \frac{p^*(y|\boldsymbol{x})p^*(\boldsymbol{x})}{p^*(y)},$$

yields that the density ratio can be expressed in terms of y as follows:

$$\begin{aligned} r^*(\boldsymbol{x}) &= \frac{p_{\mathrm{nu}}^*(\boldsymbol{x})}{p_{\mathrm{de}}^*(\boldsymbol{x})} = \left(\frac{p^*(y=+1|\boldsymbol{x})p^*(\boldsymbol{x})}{p^*(y=+1)}\right)\left(\frac{p^*(y=-1|\boldsymbol{x})p^*(\boldsymbol{x})}{p^*(y=-1)}\right)^{-1} \\ &= \frac{p^*(y=-1)}{p^*(y=+1)}\frac{p^*(y=+1|\boldsymbol{x})}{p^*(y=-1|\boldsymbol{x})}. \end{aligned}$$

The ratio $p^*(y=-1)/p^*(y=+1)$ may be simply approximated by the ratio of the number of samples:

$$\frac{p^*(y=-1)}{p^*(y=+1)} \approx \frac{n_{\mathrm{de}}/(n_{\mathrm{de}}+n_{\mathrm{nu}})}{n_{\mathrm{nu}}/(n_{\mathrm{de}}+n_{\mathrm{nu}})} = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}}.$$

The 'class' posterior probability $p^*(y|\boldsymbol{x})$ may be approximated by separating $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$ using a probabilistic classifier. Thus, given an estimator of the class posterior probability, $\widehat{p}(y|\boldsymbol{x})$, a density ratio estimator $\widehat{r}(\boldsymbol{x})$ can be constructed as

$$\widehat{r}(\boldsymbol{x}) = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}}\frac{\widehat{p}(y=+1|\boldsymbol{x})}{\widehat{p}(y=-1|\boldsymbol{x})}. \tag{6}$$

A practical advantage of the probabilistic classification approach would be its easy implementability. Indeed, one can directly use standard classification algorithms for density ratio estimation. Below, an example of probabilistic classifiers, logistic regression, is described. For making the explanation simple, we consider a set of paired samples, $\{(\boldsymbol{x}_k, y_k)\}_{k=1}^n$, where $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) := (\boldsymbol{x}_1^{\mathrm{nu}}, \ldots, \boldsymbol{x}_{n_{\mathrm{nu}}}^{\mathrm{nu}}, \boldsymbol{x}_1^{\mathrm{de}}, \ldots, \boldsymbol{x}_{n_{\mathrm{de}}}^{\mathrm{de}})$ and $(y_1, \ldots, y_n) := (+1, \ldots, +1, -1, \ldots, -1)$ for $n = n_{\mathrm{nu}} + n_{\mathrm{de}}$.

3.2 Logistic Regression Classifier

Here a popular classification algorithm called *logistic regression* (LR) (Hastie et al., 2001) is explained.

The LR classifier employs a parametric model of the following form for expressing the class-posterior probability $p^*(y|\boldsymbol{x})$,

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \left(1 + \exp\left(-y\boldsymbol{\psi}(\boldsymbol{x})^\top\boldsymbol{\theta}\right)\right)^{-1},$$

where $\boldsymbol{\psi}(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^b$ is a basis function vector and $\boldsymbol{\theta}$ ($\in \mathbb{R}^b$) is a parameter vector. The parameter vector $\boldsymbol{\theta}$ is learned so that the penalized log-likelihood is maximized, which is equivalently expressed as follows:

$$\widehat{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^m} \left[\sum_{k=1}^{n} \log\left(1 + \exp\left(-y_k\boldsymbol{\psi}(\boldsymbol{x}_k)^\top\boldsymbol{\theta}\right)\right) + \lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}\right], \tag{7}$$

where $\lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}$ is a penalty term included for regularization purposes.

Since the objective function in Eq.(7) is convex, the global optimal solution can be obtained by a standard non-linear optimization technique such as the *gradient descent method* or *(quasi-) Newton methods* (Hastie et al., 2001; Minka, 2007). An LR model classifies a new input sample $\boldsymbol{x}$ by choosing the most probable class:

$$\widehat{y} = \operatorname*{argmax}_{y=\pm1} p(y|\boldsymbol{x};\widehat{\boldsymbol{\theta}}). \tag{8}$$

Finally, a density ratio estimator $\widehat{r}_{\mathrm{LR}}(\boldsymbol{x})$ is given by

$$\begin{aligned}\widehat{r}_{\mathrm{LR}}(\boldsymbol{x}) &= \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}}\frac{1 + \exp\left(\boldsymbol{\psi}(\boldsymbol{x})^\top\widehat{\boldsymbol{\theta}}\right)}{1 + \exp\left(-\boldsymbol{\psi}(\boldsymbol{x})^\top\widehat{\boldsymbol{\theta}}\right)} = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}}\frac{\exp\left(\boldsymbol{\psi}(\boldsymbol{x})^\top\widehat{\boldsymbol{\theta}}\right)\left\{\exp\left(-\boldsymbol{\psi}(\boldsymbol{x})^\top\widehat{\boldsymbol{\theta}}\right) + 1\right\}}{1 + \exp\left(-\boldsymbol{\psi}(\boldsymbol{x})^\top\widehat{\boldsymbol{\theta}}\right)} \\ &= \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}}\exp\left(\boldsymbol{\psi}(\boldsymbol{x})^\top\widehat{\boldsymbol{\theta}}\right).\end{aligned}$$

When multi-class LR classifiers are used, density ratios among multiple densities can be estimated simultaneously. This is useful, e.g., for solving *multi-task learning* problems (Bickel et al., 2008).

When the LR model is *correctly specified*, i.e., there exists $\boldsymbol{\theta}^*$ such that $p(y|\boldsymbol{x};\boldsymbol{\theta}^*) = p^*(y|\boldsymbol{x})$, the LR approach is optimal among a class of semi-parametric estimators in the sense that the asymptotic variance is minimized (Qin, 1998). However, when the model is misspecified (which would be the case in practice), the ratio-matching approach explained in Section 4 is more preferable (Kanamori et al., 2010).

3.3 Model Selection by Cross-Validation

An important advantage of the probabilistic classification approach is that model selection (i.e., tuning the basis functions and the regularization parameter) is possible by standard CV, since the learning problem involved in this framework is a standard supervised classification problem.

More specifically, the numerator and denominator samples $D^{\mathrm{nu}} = \{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $D^{\mathrm{de}} = \{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$ are divided into T disjoint subsets $\{D_t^{\mathrm{nu}}\}_{t=1}^{T}$ and $\{D_t^{\mathrm{de}}\}_{t=1}^{T}$, respectively. Then a probabilistic classifier $\widehat{p}_t(y|\boldsymbol{x})$ is obtained using $D^{\mathrm{nu}}\backslash D_t^{\mathrm{nu}}$ and $D^{\mathrm{de}}\backslash D_t^{\mathrm{de}}$ (i.e., all samples without

D_t^{nu} and D_t^{de}), and its misclassification error (ME) for the hold-out samples D_t^{nu} and D_t^{de} is computed:

$$\widetilde{\mathrm{ME}}_t := \frac{1}{|D_t^{\mathrm{nu}}|} \sum_{\boldsymbol{x}^{\mathrm{nu}} \in D_t^{\mathrm{nu}}} I(\underset{y=\pm 1}{\mathrm{argmax}}\, \widehat{p}_t(y|\boldsymbol{x}^{\mathrm{nu}}) = +1) + \frac{1}{|D_t^{\mathrm{de}}|} \sum_{\boldsymbol{x}^{\mathrm{de}} \in D_t^{\mathrm{de}}} I(\underset{y=\pm 1}{\mathrm{argmax}}\, \widehat{p}_t(y|\boldsymbol{x}^{\mathrm{de}}) = -1),$$

where $I(\cdot)$ is the indicator function: $I(c) = 1$ if c is true and $I(c) = 0$ otherwise. This procedure is repeated for $t = 1, \ldots, T$, and the average misclassification error over all t is computed.

$$\widetilde{\mathrm{ME}} := \frac{1}{T} \sum_{t=1}^{T} \widetilde{\mathrm{ME}}_t.$$

Then the model that minimizes $\widetilde{\mathrm{ME}}$ is chosen.

3.4 Remarks

Density ratio estimation by probabilistic classification can successfully avoid density estimation by casting the problem of density ratio estimation as the problem of learning the 'class' posterior probability. An advantage of the probabilistic classification approach over the moment matching approach explained in Section 2 is that CV can be used for model selection. Furthermore, existing software packages of classification algorithms can be directly used for density ratio estimation.

As shown in Qin (1998), the probabilistic classification approach with LR has a suitable property: if the LR model is *correctly specified*, the probabilistic classification approach is optimal among a broad class of semi-parametric estimators. However, this strong theoretical property is not true when the correct model assumption is not fulfilled. In such cases, the ratio-matching approach explained in Section 4 is more preferable (Kanamori et al., 2010).

4 Ratio Matching Approach

In this section, we describe the *ratio matching* approach to density ratio estimation.

4.1 Preliminaries

A basic idea of ratio matching is to directly match a density ratio model $r(\boldsymbol{x})$ to the true density ratio $r^*(\boldsymbol{x})$ under some divergence (Figure 2). At a glance, the ratio matching problem is equivalent to the regression problem. However, ratio matching is essentially different from regression since samples of the true ratio are not available. Here, we employ the *Bregman (BR) divergence* for measuring the discrepancy between the true density ratio and the density ratio model (Bregman, 1967).

The BR divergence is an extension of the Euclidean distance to a class of distances that all share similar properties. Let f be a differentiable and strictly convex function. Then the BR divergence associated with f from t^* to t is defined as

$$\mathrm{BR}'_f(t^* \| t) := f(t^*) - f(t) - \nabla f(t)(t^* - t).$$

Note that $f(t) + \nabla f(t)(t^* - t)$ is the value of the first-order *Taylor expansion* of f around point t evaluated at point t^*. Thus, the BR divergence evaluates the difference between the value of f at point t^* and its linear extrapolation from t (Figure 3).

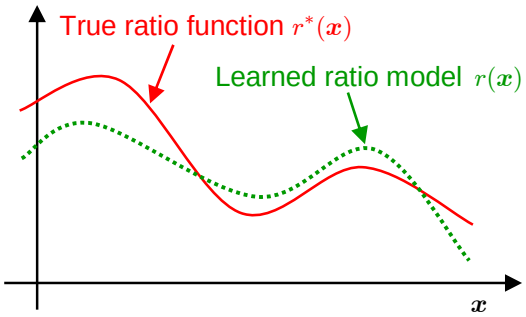

Figure 2: The idea of ratio matching.

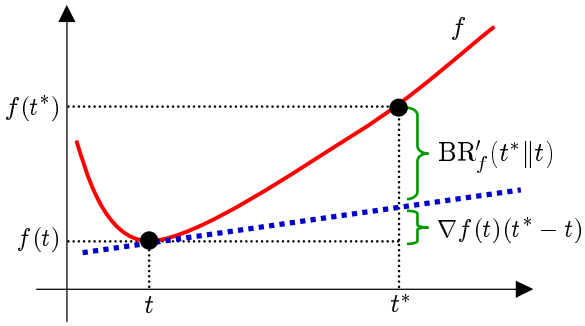

Figure 3: Bregman divergence $\mathrm{BR}'_f(t^*\|t)$.

Here the discrepancy from the true density ratio r^* to a density ratio model r is measured using the BR divergence as

$$\mathrm{BR}'_f\left(r^*\|r\right) := \int p^*_{\mathrm{de}}(\boldsymbol{x})\Big(f(r^*(\boldsymbol{x})) - f(r(\boldsymbol{x})) - \nabla f(r(\boldsymbol{x}))(r^*(\boldsymbol{x}) - r(\boldsymbol{x}))\Big)\mathrm{d}\boldsymbol{x}. \tag{9}$$

A motivation for this choice is that the BR divergence allows one to directly obtain an empirical approximation for any f. Indeed,

$$\begin{aligned}\mathrm{BR}'_f\left(r^*\|r\right) &= C - \int p^*_{\mathrm{de}}(\boldsymbol{x})f(r(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} - \int p^*_{\mathrm{de}}(\boldsymbol{x})\nabla f(r(\boldsymbol{x}))r^*(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int p^*_{\mathrm{de}}(\boldsymbol{x})\nabla f(r(\boldsymbol{x}))r(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= C + \mathrm{BR}_f\left(r\right),\end{aligned}$$

where $C := \int p^*_{\mathrm{de}}(\boldsymbol{x})f(r^*(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}$ is a constant independent of r, and

$$\mathrm{BR}_f\left(r\right) := \int p^*_{\mathrm{de}}(\boldsymbol{x})\nabla f(r(\boldsymbol{x}))r(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int p^*_{\mathrm{de}}(\boldsymbol{x})f(r(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} - \int p^*_{\mathrm{nu}}(\boldsymbol{x})\nabla f(r(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}. \tag{10}$$

Thus, an empirical approximation $\widehat{\mathrm{BR}}_f\left(r\right)$ of $\mathrm{BR}_f\left(r\right)$ is given by

$$\widehat{\mathrm{BR}}_f\left(r\right) := \frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}}\nabla f(r(\boldsymbol{x}^{\mathrm{de}}_j))r(\boldsymbol{x}^{\mathrm{de}}_j) - \frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}}f(r(\boldsymbol{x}^{\mathrm{de}}_j)) - \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}}\nabla f(r(\boldsymbol{x}^{\mathrm{nu}}_i)). \tag{11}$$

Below, ratio matching methods under the Kullback-Leibler divergence (Sugiyama et al., 2008) and the squared distance (Kanamori et al., 2009a) are explained.

4.2 Unnormalized Kullback-Leibler Divergence

In this section, a ratio matching method under the *unnormalized Kullback-Leibler (UKL) divergence* is explained.

4.2.1 Criterion

When $f(t) = t\log t - t$, the BR divergence is reduced to the UKL divergence:

$$\mathrm{UKL}'(t^*\|t) := t^*\log\frac{t^*}{t} - t^* + t.$$

Following Eqs.(10) and (11), let us denote UKL without an irrelevant constant term by $\mathrm{UKL}\,(r)$ and its empirical approximation by $\widehat{\mathrm{UKL}}\,(r)$:

$$\mathrm{UKL}\,(r) := \int p^*_{\mathrm{de}}(\boldsymbol{x}) r(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int p^*_{\mathrm{nu}}(\boldsymbol{x}) \log r(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

$$\widehat{\mathrm{UKL}}\,(r) := \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}^{\mathrm{de}}_j) - \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \log r(\boldsymbol{x}^{\mathrm{nu}}_i).$$

The density ratio model r is learned so that $\widehat{\mathrm{UKL}}(r)$ is minimized. Here, we further impose that the ratio model $r(\boldsymbol{x})$ is non-negative for all $\boldsymbol{x}$ in $\mathcal{X}$ and is normalized at $\{\boldsymbol{x}^{\mathrm{de}}_j\}_{j=1}^{n_{\mathrm{de}}}$:

$$\frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}^{\mathrm{de}}_j) = 1.$$

Then the optimization criterion is reduced to as follows.

$$\max_{r} \quad \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \log r(\boldsymbol{x}^{\mathrm{nu}}_i) \quad \text{s.t.} \quad \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}^{\mathrm{de}}_j) = 1 \quad \text{and} \quad r(\boldsymbol{x}) \geq 0 \quad \text{for all } \boldsymbol{x} \in \mathcal{X}.$$

This is called the *KL importance estimation procedure* (KLIEP). Note that the same objective function can be obtained from an empirical approximation of the KL divergence from $p^*_{\mathrm{nu}}(\boldsymbol{x})$ to $r(\boldsymbol{x})p^*_{\mathrm{de}}(\boldsymbol{x})$ (Sugiyama et al., 2008).

Below, we describe how the KLIEP formulation can be implemented for linear and kernel models. Note that the KLIEP idea can be applied to various models such as *log-linear models* (Tsuboi et al., 2009), *Gaussian mixture models* (Yamada & Sugiyama, 2009), and *mixtures of probabilistic principal component analyzers* (Yamada et al., 2010).

4.2.2 Linear and Kernel Models

For the linear density-ratio model (2), the KLIEP optimization problem is expressed as follows (Sugiyama et al., 2008):

$$\max_{\boldsymbol{\theta} \in \mathbb{R}^b} \quad \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \log(\boldsymbol{\psi}(\boldsymbol{x}^{\mathrm{nu}}_i)^\top \boldsymbol{\theta}) \quad \text{s.t.} \quad \overline{\boldsymbol{\psi}}_{\mathrm{de}}^\top \boldsymbol{\theta} = 1 \quad \text{and} \quad \boldsymbol{\theta} \geq \mathbf{0}_b,$$

where $\overline{\boldsymbol{\psi}}_{\mathrm{de}} := \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \boldsymbol{\psi}(\boldsymbol{x}^{\mathrm{de}}_j)$, and the inequality for vectors is applied in the element-wise manner. Since the above optimization problem is *convex* (i.e., the objective function to be maximized is concave and the feasible set is convex), there exists the unique global optimum solution. A pseudo code of KLIEP for linear models is described in Figure 4. As can be confirmed from the pseudo code, the denominator samples $\{\boldsymbol{x}^{\mathrm{de}}_j\}_{j=1}^{n_{\mathrm{de}}}$ appear only in terms of the basis-transformed mean $\overline{\boldsymbol{\psi}}_{\mathrm{de}}$. Thus, KLIEP is computationally very efficient even when the number n_{de} of denominator samples is very large.

4.2.3 Basis Function Design

The performance of KLIEP depends on the choice of the basis functions $\boldsymbol{\psi}(\boldsymbol{x})$. As explained below, the use of the following Gaussian kernel model would be reasonable:

$$r(\boldsymbol{x}) = \sum_{\ell=1}^{n_{\mathrm{nu}}} \theta_\ell K(\boldsymbol{x}, \boldsymbol{x}^{\mathrm{nu}}_\ell), \tag{12}$$

Input: Data samples $D^{\mathrm{nu}} = \{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $D^{\mathrm{de}} = \{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$,
and basis functions $\boldsymbol{\psi}(\boldsymbol{x})$
Output: Density ratio estimator $\widehat{r}(\boldsymbol{x})$

$\boldsymbol{\Psi}_{\mathrm{nu}} \longleftarrow (\boldsymbol{\psi}(\boldsymbol{x}_1^{\mathrm{nu}}), \ldots, \boldsymbol{\psi}(\boldsymbol{x}_{n_{\mathrm{nu}}}^{\mathrm{nu}}))^\top$;
$\overline{\boldsymbol{\psi}}_{\mathrm{de}} \longleftarrow \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \boldsymbol{\psi}(\boldsymbol{x}_j^{\mathrm{de}})$;
Initialize $\boldsymbol{\theta}$ ($> \mathbf{0}_b$) and ε ($0 < \varepsilon \ll 1$);
Repeat until convergence
$\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} + \varepsilon \boldsymbol{\Psi}_{\mathrm{nu}}^\top (\mathbf{1}_{n_{\mathrm{nu}}} ./ \boldsymbol{\Psi}_{\mathrm{nu}} \boldsymbol{\theta})$; % Gradient ascent
$\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} + (1 - \overline{\boldsymbol{\psi}}_{\mathrm{de}}^\top \boldsymbol{\theta}) \overline{\boldsymbol{\psi}}_{\mathrm{de}} / (\overline{\boldsymbol{\psi}}_{\mathrm{de}}^\top \overline{\boldsymbol{\psi}}_{\mathrm{de}})$; % Constraint satisfaction
$\boldsymbol{\theta} \longleftarrow \max(\mathbf{0}_b, \boldsymbol{\theta})$; % Constraint satisfaction
$\boldsymbol{\theta} \longleftarrow \boldsymbol{\theta} / (\overline{\boldsymbol{\psi}}_{\mathrm{de}}^\top \boldsymbol{\theta})$; % Constraint satisfaction
end
$\widehat{r}(\boldsymbol{x}) \longleftarrow \boldsymbol{\psi}(\boldsymbol{x})^\top \boldsymbol{\theta}$;

Figure 4: Pseudo code of KLIEP. './' indicates the element-wise division and $^\top$ denotes the transpose. Inequalities and the 'max' operation for vectors are applied in the element-wise manner.

Input: Data samples $D^{\mathrm{nu}} = \{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $D^{\mathrm{de}} = \{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$,
and a set of basis function candidates $\{\boldsymbol{\psi}_m(\boldsymbol{x})\}_{m=1}^{M}$
Output: Density ratio estimator $\widehat{r}(\boldsymbol{x})$

Split D^{nu} into T disjoint subsets $\{D_t^{\mathrm{nu}}\}_{t=1}^{T}$;
for each model candidate $m = 1, \ldots, M$
for each split $t = 1, \ldots, T$
$\widehat{r}_t(\boldsymbol{x}) \longleftarrow \mathrm{KLIEP}(D^{\mathrm{nu}} \backslash D_t^{\mathrm{nu}}, D^{\mathrm{de}}, \boldsymbol{\psi}(\boldsymbol{x}))$;
$\widetilde{\mathrm{UKL}}_t(m) \longleftarrow \frac{1}{|D_t^{\mathrm{nu}}|} \sum_{\boldsymbol{x} \in D_t^{\mathrm{nu}}} \log \widehat{r}_t(\boldsymbol{x})$;
end
$\widetilde{\mathrm{UKL}}(m) \longleftarrow \frac{1}{T} \sum_{t=1}^{T} \widetilde{\mathrm{UKL}}_t(m)$;
end
$\widehat{m} \longleftarrow \mathrm{argmax}_m \, \widetilde{\mathrm{UKL}}(m)$;
$\widehat{r}(\boldsymbol{x}) \longleftarrow \mathrm{KLIEP}(D^{\mathrm{nu}}, D^{\mathrm{de}}, \boldsymbol{\psi}_{\widehat{m}}(\boldsymbol{x}))$;

Figure 5: Pseudo code of CV for KLIEP.

where $K(\boldsymbol{x}, \boldsymbol{x}')$ is the Gaussian kernel (4). The reason why the numerator samples $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$, not the denominator samples $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$, are chosen as the Gaussian centers is as follows. By definition, the density ratio $r^*(\boldsymbol{x})$ tends to take large values if $p_{\mathrm{de}}^*(\boldsymbol{x})$ is small and $p_{\mathrm{nu}}^*(\boldsymbol{x})$ is large. Conversely, $r^*(\boldsymbol{x})$ tends to be small (i.e., close to zero) if $p_{\mathrm{de}}^*(\boldsymbol{x})$ is large and $p_{\mathrm{nu}}^*(\boldsymbol{x})$ is small. When a non-negative function is approximated by a Gaussian kernel model, many kernels may be needed in the region where the output of the target function is large. On the other hand, only a small number of kernels would be enough in the region where the output of the target function is close to zero (see Figure 6). Following this heuristic, many kernels are allocated in the region where $p_{\mathrm{nu}}^*(\boldsymbol{x})$ takes large values, which can be achieved by setting the Gaussian centers at $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$.

Alternatively, we may locate $(n_{\mathrm{nu}} + n_{\mathrm{de}})$ Gaussian kernels at both $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$. However, this seems not to further improve the performance, but slightly increases the computa-

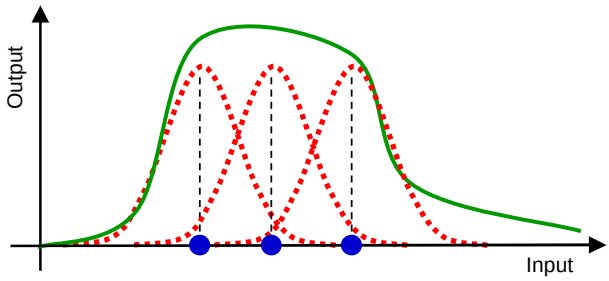

Figure 6: Heuristic of Gaussian kernel allocation.

tional cost. When n_{nu} is very large, just using all the numerator samples $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ as Gaussian centers is already computationally rather demanding. To ease this problem, a subset of $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ may be chosen in practice as Gaussian centers for computational efficiency, i.e.,

$$r(\boldsymbol{x}) = \sum_{\ell=1}^{b} \theta_\ell K(\boldsymbol{x}, \boldsymbol{c}_\ell),$$

where $\boldsymbol{c}_\ell$ is a template point randomly chosen from $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and b ($\in \{1, \ldots, n_{\mathrm{nu}}\}$) is a prefixed number.

4.2.4 Model Selection

Model selection of KLIEP is possible based on a variant of CV. More specifically, the numerator samples $D^{\mathrm{nu}} = \{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ are divided into T disjoint subsets $\{D_t^{\mathrm{nu}}\}_{t=1}^{T}$. Then a KLIEP solution $\widehat{r}_t(\boldsymbol{x})$ is obtained using $D^{\mathrm{nu}} \backslash D_t^{\mathrm{nu}}$ (i.e., all numerator samples without D_t^{nu}) and D^{de}, and its UKL value for the hold-out samples D_t^{nu} is computed:

$$\widetilde{\mathrm{UKL}}_t := \frac{1}{|D_t^{\mathrm{nu}}|} \sum_{\boldsymbol{x}^{\mathrm{nu}} \in D_t^{\mathrm{nu}}} \log \widehat{r}_t(\boldsymbol{x}^{\mathrm{nu}}).$$

This procedure is repeated for $t = 1, \ldots, T$, and the average of the above hold-out UKL values over all t is computed.

$$\widetilde{\mathrm{UKL}} := \frac{1}{T} \sum_{t=1}^{T} \widetilde{\mathrm{UKL}}_t.$$

Then the model that maximizes $\widetilde{\mathrm{UKL}}$ is chosen.

A pseudo code of CV for KLIEP is summarized in Figure 5. A MATLAB® implementation of the entire KLIEP algorithm is available from

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

4.3 Squared Distance

Here, ratio matching methods under the *squared (SQ) distance* are described.

4.3.1 Criterion

When $f(t) = \frac{1}{2}(t-1)^2$, the BR divergence is reduced to the SQ distance:

$$\mathrm{SQ}'(t^* \| t) := \frac{1}{2}(t^* - t)^2.$$

Following Eqs.(10) and (11), let us denote SQ without an irrelevant constant term by $\mathrm{SQ}(r)$ and its empirical approximation by $\widehat{\mathrm{SQ}}(r)$:

$$\mathrm{SQ}(r) := \frac{1}{2}\int p^*_{\mathrm{de}}(\boldsymbol{x}) r(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} - \int p^*_{\mathrm{nu}}(\boldsymbol{x}) r(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$
$$\widehat{\mathrm{SQ}}(r) := \frac{1}{2n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}} r(\boldsymbol{x}^{\mathrm{de}}_j)^2 - \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}} r(\boldsymbol{x}^{\mathrm{nu}}_i).$$

Here, we focus on using the linear density-ratio model (2). Since this is the same model as KLIEP for linear models, the basis design heuristic introduced in Section 4.2.3 may also be used here. For the linear density-ratio model (2), $\widehat{\mathrm{SQ}}$ is expressed as

$$\widehat{\mathrm{SQ}}(\boldsymbol{\theta}) := \frac{1}{2}\boldsymbol{\theta}^\top \widehat{\boldsymbol{H}}\boldsymbol{\theta} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\theta},$$

where

$$\widehat{\boldsymbol{H}} := \frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}} \boldsymbol{\psi}(\boldsymbol{x}^{\mathrm{de}}_j)\boldsymbol{\psi}(\boldsymbol{x}^{\mathrm{de}}_j)^\top \quad \text{and} \quad \widehat{\boldsymbol{h}} := \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}} \boldsymbol{\psi}(\boldsymbol{x}^{\mathrm{nu}}_i).$$

4.3.2 Constrained Formulation

We impose non-negativity constraint $\boldsymbol{\theta} \geq \mathbf{0}_b$ when minimizing $\widehat{\mathrm{SQ}}$. Then the optimization problem is expressed as follows.

$$\max_{\boldsymbol{\theta}\in\mathbb{R}^b} \ \frac{1}{2}\boldsymbol{\theta}^\top \widehat{\boldsymbol{H}}\boldsymbol{\theta} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\theta} + \lambda \mathbf{1}_b^\top \boldsymbol{\theta} \quad \text{s.t.} \quad \boldsymbol{\theta} \geq \mathbf{0}_b, \tag{13}$$

where λ (≥ 0) is the regularization parameter, and the constraint is imposed in order to guarantee the non-negativity of the density ratio estimator (given that the basis functions are non-negative). Together with the non-negativity constraint, the term $\mathbf{1}_b^\top \boldsymbol{\theta}$ works as the ℓ_1-regularizer:

$$\mathbf{1}_b^\top \boldsymbol{\theta} = \|\boldsymbol{\theta}\|_1 := \sum_{\ell=1}^{b} |\theta_\ell|.$$

This formulation is called *least-squares importance fitting* (LSIF) (Kanamori et al., 2009a). The LSIF optimization problem is a convex quadratic program. Therefore, the unique global optimal solution can be computed by a standard optimization package.

We can also use the ℓ_2-regularizer $\boldsymbol{\theta}^\top\boldsymbol{\theta}$, instead of the ℓ_1-regularizer $\mathbf{1}_b^\top\boldsymbol{\theta}$, without changing the computational property. However, using the ℓ_1-regularizer would be more advantageous since the solution tends to be *sparse* (Williams, 1995; Tibshirani, 1996; Chen et al., 1998). Furthermore, as explained in Section 4.3.3, the use of the ℓ_1-regularizer allows one to compute the entire *regularization path* efficiently.

Model selection of LSIF (i.e., the choice of the basis functions and the regularization parameter) is possible by CV based on the SQ distance. More specifically, the numerator and

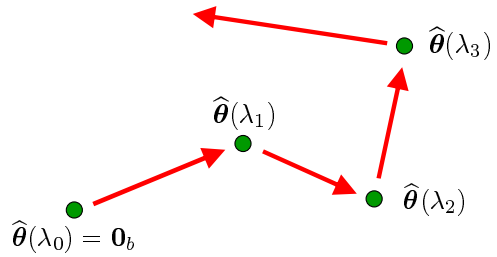

Figure 7: Regularization path tracking of LSIF. The solution $\widehat{\boldsymbol{\theta}}(\lambda)$ is shown to be piecewise-linear in the parameter space as a function of λ. Starting from $\lambda = \infty$, the trajectory of the solution is traced as λ is decreased to zero. When $\lambda \geq \lambda_0$ for some $\lambda_0 \geq 0$, the solution stays at the origin $\mathbf{0}_b$. When λ gets smaller than λ_0, the solution departs from the origin. As λ is further decreased, for some λ_1 such that $0 \leq \lambda_1 \leq \lambda_0$, the solution goes straight to $\widehat{\boldsymbol{\theta}}(\lambda_1)$ with a constant 'speed'. Then the solution path changes the direction and, for some λ_2 such that $0 \leq \lambda_2 \leq \lambda_1$, the solution is headed straight for $\widehat{\boldsymbol{\theta}}(\lambda_2)$ with a constant speed as λ is further decreased. This process is repeated until λ reaches zero.

denominator samples $D^{\mathrm{nu}} = \{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $D^{\mathrm{de}} = \{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$ are divided into T disjoint subsets $\{D_t^{\mathrm{nu}}\}_{t=1}^T$ and $\{D_t^{\mathrm{de}}\}_{t=1}^T$, respectively. Then a density ratio estimator $\widehat{r}_t(\boldsymbol{x})$ is obtained using $D^{\mathrm{nu}} \backslash D_t^{\mathrm{nu}}$ and $D^{\mathrm{de}} \backslash D_t^{\mathrm{de}}$ (i.e., all samples without D_t^{nu} and D_t^{de}), and its SQ value for the hold-out samples D_t^{nu} and D_t^{de} is computed:

$$\widetilde{\mathrm{SQ}}_t := \frac{1}{2|D_t^{\mathrm{nu}}|} \sum_{\boldsymbol{x}^{\mathrm{nu}} \in D_t^{\mathrm{nu}}} r(\boldsymbol{x}^{\mathrm{nu}})^2 - \frac{1}{|D_t^{\mathrm{de}}|} \sum_{\boldsymbol{x}^{\mathrm{de}} \in D_t^{\mathrm{de}}} r(\boldsymbol{x}^{\mathrm{de}}).$$

This procedure is repeated for $t = 1, \ldots, T$, and the average of the above hold-out SQ values is computed.

$$\widetilde{\mathrm{SQ}} := \frac{1}{T} \sum_{t=1}^{T} \widetilde{\mathrm{SQ}}_t.$$

Then the model that minimizes $\widetilde{\mathrm{SQ}}$ is chosen.

For LSIF, an *information criterion* (Akaike, 1974) is also available for model selection (Kanamori et al., 2009a).

4.3.3 Entire Regularization Path

The LSIF solution $\widehat{\boldsymbol{\theta}}$ is shown to be *piecewise-linear* with respect to the regularization parameter λ (see Figure 7). Thus, the *regularization path* (i.e., solutions for all λ) can be computed efficiently based on the *parametric optimization technique* (Best, 1982; Efron et al., 2004; Hastie et al., 2004).

A basic idea of regularization path tracking is to check the violation of the *Karush-Kuhn-Tacker (KKT) conditions* (Boyd & Vandenberghe, 2004)—which are necessary and sufficient for optimality of convex programs—when the regularization parameter λ is changed. A pseudo code of the regularization path tracking algorithm for LSIF is described in Figure 8. Thanks to the regularization path algorithm, LSIF is computationally efficient in model selection scenarios, where solutions for various λ are computed.

The pseudo code implies that we no longer need a quadratic programming solver for obtaining the solution of LSIF—just computing matrix inverses is sufficient. Furthermore, the regularization path algorithm is computationally more efficient when the solution is sparse, that is, most

Input: $\widehat{\boldsymbol{H}}$ and $\widehat{\boldsymbol{h}}$
Output: entire regularization path $\widehat{\boldsymbol{\theta}}(\lambda)$ for $\lambda \geq 0$

$\tau \longleftarrow 0$; $\quad k \longleftarrow \mathrm{argmax}_i\{\widehat{h}_i \mid i = 1, \ldots, b\}$; $\quad \lambda_\tau \longleftarrow \widehat{h}_k$;
$\widehat{\mathcal{A}} \longleftarrow \{1, \ldots, b\} \backslash \{k\}$; $\quad \widehat{\boldsymbol{\theta}}(\lambda_\tau) \longleftarrow \mathbf{0}_b$;
While $\lambda_\tau > 0$
$\quad \widehat{\boldsymbol{E}} \longleftarrow \boldsymbol{O}_{|\widehat{\mathcal{A}}| \times b}$;
$\quad$ **For** $i = 1, \ldots, |\widehat{\mathcal{A}}|$
$\quad\quad \widehat{E}_{i,\widehat{j}_i} \longleftarrow 1$; $\%$ $\widehat{\mathcal{A}} = \{\widehat{j}_1, \ldots, \widehat{j}_{|\widehat{\mathcal{A}}|} \mid \widehat{j}_1 < \cdots < \widehat{j}_{|\widehat{\mathcal{A}}|}\}$
$\quad$ **end**
$\quad \widehat{\boldsymbol{G}} \longleftarrow \begin{pmatrix} \widehat{\boldsymbol{H}} & -\widehat{\boldsymbol{E}}^\top \\ -\widehat{\boldsymbol{E}} & \boldsymbol{O}_{|\widehat{\mathcal{A}}| \times |\widehat{\mathcal{A}}|} \end{pmatrix}$; $\quad \boldsymbol{u} \longleftarrow \widehat{\boldsymbol{G}}^{-1} \begin{pmatrix} \widehat{\boldsymbol{h}} \\ \mathbf{0}_{|\widehat{\mathcal{A}}|} \end{pmatrix}$; $\quad \boldsymbol{v} \longleftarrow \widehat{\boldsymbol{G}}^{-1} \begin{pmatrix} \mathbf{1}_b \\ \mathbf{0}_{|\widehat{\mathcal{A}}|} \end{pmatrix}$;
$\quad$ **If** $\boldsymbol{v} \leq \mathbf{0}_{b+|\widehat{\mathcal{A}}|}$ $\%$ the final interval
$\quad\quad \lambda_{\tau+1} \longleftarrow 0$; $\quad \widehat{\boldsymbol{\theta}}(\lambda_{\tau+1}) \longleftarrow (u_1, \ldots, u_b)^\top$;
$\quad$ **else** $\%$ an intermediate interval
$\quad\quad k \longleftarrow \mathrm{argmax}_i\{u_i / v_i \mid v_i > 0,\ i = 1, \ldots, b + |\widehat{\mathcal{A}}|\}$; $\quad \lambda_{\tau+1} \longleftarrow \max\{0, u_k / v_k\}$;
$\quad\quad \widehat{\boldsymbol{\theta}}(\lambda_{\tau+1}) \longleftarrow (u_1, \ldots, u_b)^\top - \lambda_{\tau+1}(v_1, \ldots, v_b)^\top$;
$\quad\quad$ **If** $1 \leq k \leq b$
$\quad\quad\quad \widehat{\mathcal{A}} \longleftarrow \widehat{\mathcal{A}} \cup \{k\}$;
$\quad\quad$ **else**
$\quad\quad\quad \widehat{\mathcal{A}} \longleftarrow \widehat{\mathcal{A}} \backslash \{\widehat{j}_{k-b}\}$;
$\quad\quad$ **end**
$\quad$ **end**
$\quad \tau \longleftarrow \tau + 1$;
end

$$\widehat{\boldsymbol{\theta}}(\lambda) \longleftarrow \begin{cases} \mathbf{0}_b & \text{if } \lambda \geq \lambda_0 \\ \frac{\lambda_{\tau+1} - \lambda}{\lambda_{\tau+1} - \lambda_\tau} \widehat{\boldsymbol{\theta}}(\lambda_\tau) + \frac{\lambda - \lambda_\tau}{\lambda_{\tau+1} - \lambda_\tau} \widehat{\boldsymbol{\theta}}(\lambda_{\tau+1}) & \text{if } \lambda_{\tau+1} \leq \lambda \leq \lambda_\tau \end{cases}$$

Figure 8: Pseudo code for computing the entire regularization path of LSIF. When the computation of $\widehat{\boldsymbol{G}}^{-1}$ is numerically unstable, we may add small positive diagonals to $\widehat{\boldsymbol{H}}$ for stabilization purposes.

of the elements are zero since the number of change points tends to be small for such sparse solutions. However, the regularization path tracking algorithm was found to be numerically rather unstable (Kanamori et al., 2009a).

An R implementation of LSIF is available from

`http://www.math.cm.is.nagoya-u.ac.jp/~kanamori/software/LSIF/`

4.3.4 Unconstrained Formulation

The regularization path tracking algorithm for LSIF was shown to suffer from a numerical problem, and therefore is not practically reliable. Here, a practical alternative to LSIF is introduced, which gives an approximate solution to LSIF in a computationally efficient and reliable manner (Kanamori et al., 2009a).

The approximation idea introduced here is very simple: the non-negativity constraint of the

parameters in the optimization problem (13) is ignored. This results in the following unconstrained optimization problem.

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^b}\left[\frac{1}{2}\boldsymbol{\beta}^\top\widehat{\boldsymbol{H}}\boldsymbol{\beta}-\widehat{\boldsymbol{h}}^\top\boldsymbol{\beta}+\frac{\lambda}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right]. \tag{14}$$

In the above, a quadratic regularization term $\boldsymbol{\beta}^\top\boldsymbol{\beta}/2$ was included instead of the linear one $\mathbf{1}_b^\top\boldsymbol{\theta}$ since the linear penalty term does not work as a regularizer without the non-negativity constraint. Eq.(14) is an unconstrained convex quadratic program, and the solution can be analytically computed as

$$\widetilde{\boldsymbol{\beta}}=(\widehat{\boldsymbol{H}}+\lambda\boldsymbol{I}_b)^{-1}\widehat{\boldsymbol{h}},$$

where $\boldsymbol{I}_b$ is the b-dimensional identity matrix. Since the non-negativity constraint $\boldsymbol{\beta}\geq\mathbf{0}_b$ was dropped, some of the learned parameters could be negative. To compensate for this approximation error, the solution is modified as

$$\widehat{\boldsymbol{\beta}}=\max(\mathbf{0}_b,\widetilde{\boldsymbol{\beta}}),$$

where the 'max' operation for a pair of vectors is applied in the element-wise manner. This is the solution of the approximation method called *unconstrained LSIF* (uLSIF) (Kanamori et al., 2009a). An advantage of uLSIF is that the solution can be computed just by solving a system of linear equations. Therefore, its computation is stable when λ is not too small.

Due to the ℓ_2-regularizer, the solution tends to be close to $\mathbf{0}_b$ to some extent. Thus, the effect of ignoring the non-negativity constraint may not be so critical See Kanamori et al. (2009a) for theoretical and experimental error analysis.

4.3.5 Analytic Expression of Leave-One-Out Score

A practically important advantage of uLSIF over LSIF is that the score of *leave-one-out CV* (LOOCV) can be computed analytically (Kanamori et al., 2009a)—thanks to this property, the computational complexity for performing LOOCV is the same order as just computing a single solution.

In the current setup, two sets of samples, $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$, generally have different sample size. For simplicity, we assume that $n_{\mathrm{de}}\leq n_{\mathrm{nu}}$ and the i-th numerator sample $\boldsymbol{x}_i^{\mathrm{nu}}$ and the i-th denominator sample $\boldsymbol{x}_i^{\mathrm{de}}$ are held out at the same time; the numerator samples $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=n_{\mathrm{de}}+1}^{n_{\mathrm{nu}}}$ are always used for density ratio estimation. Note that this assumption is only for the sake of simplicity; the order of numerator samples can be changed without sacrificing the computational advantages.

Let $\widehat{r}^{(i)}(\boldsymbol{x})$ be a density ratio estimate obtained without the i-th numerator sample $\boldsymbol{x}_i^{\mathrm{nu}}$ and the i-th denominator sample $\boldsymbol{x}_i^{\mathrm{de}}$. Then the LOOCV score is expressed as

$$\mathrm{LOOCV}=\frac{1}{n_{\mathrm{de}}}\sum_{i=1}^{n_{\mathrm{de}}}\left[\frac{1}{2}(\widehat{r}^{(i)}(\boldsymbol{x}_i^{\mathrm{de}}))^2-\widehat{r}^{(i)}(\boldsymbol{x}_i^{\mathrm{nu}})\right].$$

A trick to efficiently compute the LOOCV score is to use the *Sherman-Woodbury-Morrison formula* (Golub & Loan, 1996) for computing matrix inverses: for an invertible square matrix $\boldsymbol{A}$ and vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ such that $\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}\neq-1$,

$$(\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^\top)^{-1}=\boldsymbol{A}^{-1}-\frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top\boldsymbol{A}^{-1}}{1+\boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}}.$$

Input: $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $\{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$
Output: $\widehat{r}(\boldsymbol{x})$

$b \longleftarrow \min(100, n_{\mathrm{nu}})$; $\quad n \longleftarrow \min(n_{\mathrm{nu}}, n_{\mathrm{de}})$;
Randomly choose b centers $\{\boldsymbol{c}_\ell\}_{\ell=1}^{b}$ from $\{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ without replacement;
For each candidate of Gaussian width σ

$$\widehat{H}_{\ell,\ell'} \longleftarrow \frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}}\exp\left(-\frac{\|\boldsymbol{x}_j^{\mathrm{de}}-\boldsymbol{c}_\ell\|^2+\|\boldsymbol{x}_j^{\mathrm{de}}-\boldsymbol{c}_{\ell'}\|^2}{2\sigma^2}\right) \text{ for } \ell,\ell'=1,\ldots,b;$$

$$\widehat{h}_\ell \longleftarrow \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}}\exp\left(-\frac{\|\boldsymbol{x}_i^{\mathrm{nu}}-\boldsymbol{c}_\ell\|^2}{2\sigma^2}\right) \text{ for } \ell=1,\ldots,b;$$

$$X_{\ell,i}^{\mathrm{nu}} \longleftarrow \exp\left(-\frac{\|\boldsymbol{x}_i^{\mathrm{nu}}-\boldsymbol{c}_\ell\|^2}{2\sigma^2}\right) \text{ for } i=1,\ldots,n \text{ and } \ell=1,\ldots,b;$$

$$X_{\ell,i}^{\mathrm{de}} \longleftarrow \exp\left(-\frac{\|\boldsymbol{x}_i^{\mathrm{de}}-\boldsymbol{c}_\ell\|^2}{2\sigma^2}\right) \text{ for } i=1,\ldots,n \text{ and } \ell=1,\ldots,b;$$

For each candidate of regularization parameter λ

$$\widehat{\boldsymbol{B}} \longleftarrow \widehat{\boldsymbol{H}} + \frac{\lambda(n_{\mathrm{de}}-1)}{n_{\mathrm{de}}}\boldsymbol{I}_b;$$

$$\boldsymbol{B}_0 \longleftarrow \widehat{\boldsymbol{B}}^{-1}\widehat{\boldsymbol{h}}\boldsymbol{1}_n^\top + \widehat{\boldsymbol{B}}^{-1}\boldsymbol{X}^{\mathrm{de}}\,\mathrm{diag}\left(\frac{\widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{B}}^{-1}\boldsymbol{X}^{\mathrm{de}}}{n_{\mathrm{de}}\boldsymbol{1}_n^\top - \boldsymbol{1}_b^\top(\boldsymbol{X}^{\mathrm{de}} * \widehat{\boldsymbol{B}}^{-1}\boldsymbol{X}^{\mathrm{de}})}\right);$$

$$\boldsymbol{B}_1 \longleftarrow \widehat{\boldsymbol{B}}^{-1}\boldsymbol{X}^{\mathrm{nu}} + \widehat{\boldsymbol{B}}^{-1}\boldsymbol{X}^{\mathrm{de}}\,\mathrm{diag}\left(\frac{\boldsymbol{1}_b^\top(\boldsymbol{X}^{\mathrm{nu}} * \widehat{\boldsymbol{B}}^{-1}\boldsymbol{X}^{\mathrm{de}})}{n_{\mathrm{de}}\boldsymbol{1}_n^\top - \boldsymbol{1}_b^\top(\boldsymbol{X}^{\mathrm{de}} * \widehat{\boldsymbol{B}}^{-1}\boldsymbol{X}^{\mathrm{de}})}\right);$$

$$\boldsymbol{B}_2 \longleftarrow \max\left(\boldsymbol{O}_{b\times n},\ \frac{n_{\mathrm{de}}-1}{n_{\mathrm{de}}(n_{\mathrm{nu}}-1)}(n_{\mathrm{nu}}\boldsymbol{B}_0 - \boldsymbol{B}_1)\right);$$

$$\boldsymbol{w}_{\mathrm{de}} \longleftarrow (\boldsymbol{1}_b^\top(\boldsymbol{X}^{\mathrm{de}} * \boldsymbol{B}_2))^\top; \quad \boldsymbol{w}_{\mathrm{nu}} \longleftarrow (\boldsymbol{1}_b^\top(\boldsymbol{X}^{\mathrm{nu}} * \boldsymbol{B}_2))^\top;$$

$$\mathrm{LOOCV}(\sigma,\lambda) \longleftarrow \frac{\boldsymbol{w}_{\mathrm{de}}^\top\boldsymbol{w}_{\mathrm{de}}}{2n} - \frac{\boldsymbol{1}_n^\top\boldsymbol{w}_{\mathrm{nu}}}{n};$$

end
end

$$(\widehat{\sigma},\widehat{\lambda}) \longleftarrow \mathrm{argmin}_{(\sigma,\lambda)}\,\mathrm{LOOCV}(\sigma,\lambda);$$

$$\widetilde{H}_{\ell,\ell'} \longleftarrow \frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}}\exp\left(-\frac{\|\boldsymbol{x}_j^{\mathrm{de}}-\boldsymbol{c}_\ell\|^2+\|\boldsymbol{x}_j^{\mathrm{de}}-\boldsymbol{c}_{\ell'}\|^2}{2\widehat{\sigma}^2}\right) \text{ for } \ell,\ell'=1,\ldots,b;$$

$$\widetilde{h}_\ell \longleftarrow \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}}\exp\left(-\frac{\|\boldsymbol{x}_i^{\mathrm{nu}}-\boldsymbol{c}_\ell\|^2}{2\widehat{\sigma}^2}\right) \text{ for } \ell=1,\ldots,b;$$

$$\widehat{\boldsymbol{\alpha}} \longleftarrow \max(\boldsymbol{0}_b, (\widetilde{\boldsymbol{H}} + \widehat{\lambda}\boldsymbol{I}_b)^{-1}\widetilde{\boldsymbol{h}});$$

$$\widehat{w}(\boldsymbol{x}) \longleftarrow \sum_{\ell=1}^{b}\widehat{\alpha}_\ell\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}_\ell\|^2}{2\widehat{\sigma}^2}\right);$$

Figure 9: Pseudo code of uLSIF algorithm with LOOCV. $\boldsymbol{B} * \boldsymbol{B}'$ denotes the element-wise multiplication of matrices $\boldsymbol{B}$ and $\boldsymbol{B}'$ of the same size, that is, the (i,j)-th element is given by $B_{i,j}B'_{i,j}$. For n-dimensional vectors $\boldsymbol{b}$ and $\boldsymbol{b}'$, $\mathrm{diag}\left(\frac{\boldsymbol{b}}{\boldsymbol{b}'}\right)$ denotes the $n\times n$ diagonal matrix with i-th diagonal element b_i/b'_i.

A pseudo code of uLSIF with LOOCV-based model selection is summarized in Figure 9. Note that the basis design heuristic explained in Section 4.2.3 is used in the pseudo code, but the analytic form of the LOOCV score is available for any basis functions.

A MATLAB® implementation of uLSIF is available from

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

and an R implementation of uLSIF is available from

`http://www.math.cm.is.nagoya-u.ac.jp/~kanamori/software/LSIF/`

4.4 Remarks

Density ratio estimation by ratio matching can successfully avoid density estimation. Furthermore, CV based on the target divergence functional is available for model selection.

We have described the ratio matching methods for the UKL divergence and the SQ distance. The UKL method (KLIEP) is applicable to a variety of models such as linear/kernel models (Sugiyama et al., 2008), log-linear models (Tsuboi et al., 2009), mixtures of Gaussians (Yamada & Sugiyama, 2009), and mixtures of probabilistic principal component analyzers (Yamada et al., 2010). On the other hand, the SQ methods are computationally more efficient. The constrained method (LSIF) for the ℓ_1-regularizer is equipped with a regularization path tracking algorithm. Furthermore, its unconstrained variant (uLSIF) allows one to compute the density ratio estimator *analytically*; the leave-one-out CV score can also be computed in a closed form. Thus, the overall computation of uLSIF including model selection is highly efficient (Kanamori et al., 2009a).

The fact that uLSIF has an analytic-form solution is actually very useful beyond its computational efficiency. When one wants to optimize some criterion defined using a density ratio estimate (e.g., *mutual information*, Cover & Thomas, 2006), the analytic-form solution of uLSIF allows one to compute the *derivative* of the target criterion analytically. Then one can develop, e.g., gradient-based algorithms and (quasi-) Newton algorithms for optimization. This property can be successfully utilized, e.g., in identifying the central subspace in *sufficient dimensionality reduction* (Suzuki & Sugiyama, 2010), finding independent components in *independent component analysis* (Suzuki & Sugiyama, 2009), performing dependence minimizing regression in *causal inference* (Yamada & Sugiyama, 2010), and identifying the hetero-distributional subspace in *direct density ratio estimation with dimensionality reduction* (Sugiyama et al., 2010a).

The ratio matching approach can also be characterized as divergence estimation. Let f be a convex function such that $f(1) = 0$. Then the *Ali-Silver-Csiszár (ASC) divergence* associated with f from p^*_{de} to p^*_{nu} is defined as follows (Ali & Silvey, 1966; Csiszár, 1967):

$$\mathrm{ASC}_f(p^*_{\mathrm{de}} \| p^*_{\mathrm{nu}}) := \int p^*_{\mathrm{de}}(\boldsymbol{x}) f\left(\frac{p^*_{\mathrm{nu}}(\boldsymbol{x})}{p^*_{\mathrm{de}}(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}.$$

Let f^* be the *Legendre-Fenchel dual* of f (Rockafellar, 1970):

$$f^*(s) := \sup_{t^*} \left[t^* s - f(t^*)\right].$$

The convexity of f implies

$$f(t^*) \geq f(t) + (t^* - t)\nabla f(t),$$

where the equality holds if and only if $t^* = t$ (see Figure 3 again). Thus, for $s = \nabla f(t)$, $f^*(s)$ is expressed as

$$\begin{aligned} f^*(\nabla f(t)) &= \sup_{t^*} [t^* \nabla f(t) - f(t^*)] \\ &= t \nabla f(t) - f(t). \end{aligned}$$

Then, the BR divergence associated with f from r^* to r without an irrelevant constant (see Eq.(10)) can be expressed in terms of f^* as

$$\mathrm{BR}_f\left(r^* \| r\right) = \int p^*_{\mathrm{de}}(\boldsymbol{x}) \Big(f^*(\nabla f(r(\boldsymbol{x}))) - \nabla f(r(\boldsymbol{x})) r^*(\boldsymbol{x}) \Big) \mathrm{d}\boldsymbol{x}. \tag{15}$$

Eq.(15) is minimized with respect to r if and only if $r = r^*$ (Nguyen et al., 2010):

$$\min_r \mathrm{BR}_f\left(r^* \| r\right) = \int p^*_{\mathrm{de}}(\boldsymbol{x}) f(r^*(\boldsymbol{x})) \mathrm{d}\boldsymbol{x}.$$

Consequently, the ASC divergence can be approximated as

$$\begin{aligned} \mathrm{ASC}_f(p^*_{\mathrm{de}} \| p^*_{\mathrm{nu}}) &= \min_r \mathrm{BR}_f\left(r^* \| r\right) \\ &\approx \min_r \left[\frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} f^*(\nabla f(r(\boldsymbol{x}^{\mathrm{de}}_j))) - \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \nabla f(r(\boldsymbol{x}^{\mathrm{nu}}_i)) \right]. \end{aligned}$$

This agrees with the ASC-divergence estimator proposed in Nguyen et al. (2010).

5 Conclusions

In this paper, we provided a comprehensive review of density ratio estimation methods, including the moment matching approach (Section 2), the probabilistic classification approach (Section 3), and the ratio matching approach (Section 4). Through extensive experiments, these methods were shown to outperform the naive approach of taking the ratio of kernel density estimators (Sugiyama et al., 2008; Gretton et al., 2009; Kanamori et al., 2009a; Hido et al., 2010).

Theoretical analysis of these direct density-ratio estimators has also been carried out. For example, in Kanamori et al. (2010), the accuracy of (A) the ratio of maximum likelihood density estimators, (B) probabilistic classification with logistic regression, and (C) ratio matching under the Kullback-Leibler divergence has been theoretically compared in the parametric setup. The paper showed that, when the numerator and denominator densities are known to be members of the exponential family, (A) is better than (B) and (B) is better than (C). On the other hand, once the model assumption is violated, (C) was shown to be better than (A) and (B). Thus, in practical situations where no exact model is available, (C) would be the most promising approach to density ratio estimation.

For non-parametric cases, the convergence rate of the infinite-order moment matching approach (Gretton et al., 2009), ratio matching under the Kullback-Leibler divergence (Sugiyama et al., 2008; Nguyen et al., 2010), and ratio matching under the squared distance (Kanamori et al., 2009b) has been elucidated. However, it seems to be an open research topic to theoretically prove that these direct density-ratio estimators are really superior to the naive approach of taking the ratio of non-parametric density estimators.

Finally, the performance of density ratio estimation in high-dimensional problems can be further improved by *dimensionality reduction*. More specifically, density ratio estimation is

carried out only in a subspace in which the numerator and denominator densities are significantly different. Such approaches have been explored recently (Sugiyama et al., 2010b; Sugiyama et al., 2010a), and would be a promising direction for further improving the estimation accuracy of density ratios.

Acknowledgment

MS was supported by SCAT, AOARD, and the JST PRESTO program. TS was supported by Global COE Program "The research and training center for new development in mathematics", MEXT, Japan. TK was supported by MEXT Grant-in-Aid for Young Scientists 20700251.

References

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

Ali, S. M., & Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, *28*, 131–142.

Best, M. J. (1982). *An algorithm for the solution of the parametric quadratic programming problem* (Technical Report 82-24). Faculty of Mathematics, University of Waterloo.

Bickel, S., Bogojeska, J., Lengauer, T., & Scheffer, T. (2008). Multi-task learning for HIV therapy screening. *Proceedings of 25th Annual International Conference on Machine Learning (ICML2008)* (pp. 56–63). Helsinki, Finland: Omnipress.

Bickel, S., Brückner, M., & Scheffer, T. (2007). Discriminative learning for differing training and test distributions. *Proceedings of the 24th International Conference on Machine Learning* (pp. 81–88).

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge, UK: Cambridge University Press.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, *7*, 200–217.

Chen, S. S., Donoho, D. L., & Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, *20*, 33–61.

Cheng, K. F., & Chu, C. K. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, *10*, 583–604.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. Hoboken, NJ, USA: John Wiley & Sons, Inc. 2nd edition.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, *2*, 229–318.

Efron, B., Hastie, T., Tibshirani, R., & Johnstone, I. (2004). Least angle regression. *The Annals of Statistics*, *32*, 407–499.

Golub, G. H., & Loan, C. F. V. (1996). *Matrix computations*. Baltimore, MD: Johns Hopkins University Press.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., & Schölkopf, B. (2009). Covariate shift by kernel mean matching. In J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer and N. Lawrence (Eds.), *Dataset shift in machine learning*, chapter 8, 131–160. Cambridge, MA: MIT Press.

Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, *5*, 1391–1415.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., & Kanamori, T. (2008). Inlier-based outlier detection via direct density ratio estimation. *Proceedings of IEEE International Conference on Data Mining (ICDM2008)* (pp. 223–232). Pisa, Italy.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., & Kanamori, T. (2010). Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*. to appear.

Kanamori, T., Hido, S., & Sugiyama, M. (2009a). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, *10*, 1391–1445.

Kanamori, T., Suzuki, T., & Sugiyama, M. (2009b). *Condition number analysis of kernel-based density ratio estimation* (Technical Report). arXiv. http://www.citebase.org/abstract?id=oai:arXiv.org:0912.2800

Kanamori, T., Suzuki, T., & Sugiyama, M. (2010). Theoretical analysis of density ratio estimation. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*. to appear.

Kawahara, Y., & Sugiyama, M. (2009). Change-point detection in time-series data by direct density-ratio estimation. *Proceedings of 2009 SIAM International Conference on Data Mining (SDM2009)* (pp. 389–400). Sparks, Nevada, USA.

Minka, T. P. (2007). *A comparison of numerical optimizers for logistic regression* (Technical Report). Microsoft Research. http://research.microsoft.com/˜minka/papers/logreg/minka-logreg.pdf

Nguyen, X., Wainwright, M. J., & Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*. to appear.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, *85*, 619–639.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. (Eds.). (2009). *Dataset shift in machine learning*. Cambridge, MA: MIT Press.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton, NJ, USA: Princeton University Press.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

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

Smola, A., Song, L., & Teo, C. H. (2009). Relative novelty detection. *Twelfth International Conference on Artificial Intelligence and Statistics* (pp. 536–543).

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, *2*, 67–93.

Sugiyama, M. (2010). Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. *IEICE Transactions on Information and Systems*. submitted.

Sugiyama, M., Hara, S., von Bünau, P., Suzuki, T., Kanamori, T., & Kawanabe, M. (2010a). Direct density ratio estimation with dimensionality reduction. *Proceedings of 2010 SIAM International Conference on Data Mining (SDM2010)*. Columbus, Ohio, USA.

Sugiyama, M., Kanamori, T., Suzuki, T., Hido, S., Sese, J., Takeuchi, I., & Wang, L. (2009). A density-ratio framework for statistical data processing. *IPSJ Transactions on Computer Vision and Applications*, *1*, 183–208.

Sugiyama, M., Kawanabe, M., & Chui, P. L. (2010b). Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, *23*, 44–59.

Sugiyama, M., Krauledat, M., & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, *8*, 985–1005.

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

Sugiyama, M., Suzuki, T., & Kanamori, T. (2011). *Density ratio estimation in machine learning: A versatile tool for statistical data processing*. Cambridge, UK: Cambridge University Press. to appear.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., & Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, *60*, 699–746.

Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H., & Okanohara, D. (2010c). Least-squares conditional density estimation. *IEICE Transactions on Information and Systems*, *E93-D*, 583–594.

Sugiyama, M., von Bünau, P., Kawanabe, M., & Müller, K.-R. (2010d). *Covariate shift adaptation: Towards machine learning in non-stationary environment*. Cambridge, MA: MIT Press. to appear.

Suzuki, T., & Sugiyama, M. (2009). Estimating squared-loss mutual information for independent component analysis. *Independent Component Analysis and Signal Separation* (pp. 130–137). Berlin: Springer.

Suzuki, T., & Sugiyama, M. (2010). Sufficient dimension reduction via squared-loss mutual information estimation. *Proceedings of The Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010)* (pp. 804–811). Chia Laguna, Sardinia, Italy.

Suzuki, T., Sugiyama, M., Kanamori, T., & Sese, J. (2009a). Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, *10*, S52.

Suzuki, T., Sugiyama, M., Sese, J., & Kanamori, T. (2008). Approximating mutual information by maximum likelihood density ratio estimation. *JMLR Workshop and Conference Proceedings* (pp. 5–20).

Suzuki, T., Sugiyama, M., & Tanaka, T. (2009b). Mutual information approximation via maximum likelihood estimation of density ratio. *Proceedings of 2009 IEEE International Symposium on Information Theory (ISIT2009)* (pp. 463–467). Seoul, Korea.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., & Sugiyama, M. (2009). Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, *17*, 138–155.

Williams, P. M. (1995). Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, *7*, 117–143.

Yamada, M., & Sugiyama, M. (2009). Direct importance estimation with Gaussian mixture models. *IEICE Transactions on Information and Systems*, *E92-D*, 2159–2162.

Yamada, M., & Sugiyama, M. (2010). Dependence minimizing regression with model selection for non-linear causal inference under non-Gaussian noise. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI2010)*. to appear.

Yamada, M., Sugiyama, M., Wichern, G., & Simm, J. (2010). Direct importance estimation with a mixture of probabilistic principal component analyzers. *IEICE Transactions on Information and Systems*. submitted.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. *Proceedings of the Twenty-First International Conference on Machine Learning* (pp. 903–910). New York, NY: ACM Press.