In B. Schölkopf, Z. Luo, and V. Vovk (Eds.),
Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik,
Chapter 23, pp.273–283, Springer, Berlin, Germany, 2013.

Direct Approximation of Divergences between Probability Distributions

Masashi Sugiyama
Tokyo Institute of Technology, Japan.
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/˜sugi

Abstract

Approximating a divergence between two probability distributions from their samples is a fundamental challenge in the statistics, information theory, and machine learning communities, because a divergence estimator can be used for various purposes such as two-sample homogeneity testing, change-point detection, and class-balance estimation. Furthermore, an approximator of a divergence between the joint distribution and the product of marginals can be used for independence testing, which has a wide range of applications including feature selection and extraction, clustering, object matching, independent component analysis, and causality learning. In this article, we review recent advances in direct divergence approximation that follow the general inference principle advocated by Vladimir Vapnik—one should not solve a more general problem as an intermediate step. More specifically, direct divergence approximation avoids separately estimating two probability distributions when approximating a divergence. We cover direct approximators of the Kullback-Leibler (KL) divergence, the Pearson (PE) divergence, the relative PE (rPE) divergence, and the L^2-distance. Despite the overwhelming popularity of the KL divergence, we argue that the latter approximators are more useful in practice due to their computational efficiency, high numerical stability, and superior robustness against outliers.

1 Introduction

Let us consider the problem of approximating a divergence D between two probability distributions P and P' on $\mathbb{R}^d$ from two sets of independent and identically distributed samples $\mathcal{X} := \{\boldsymbol{x}_i\}_{i=1}^n$ and $\mathcal{X}' := \{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ following P and P'.

A divergence approximator can be used for various purposes such as two-sample testing [25, 11], change detection in time-series [13], class-prior estimation under class-balance change [20], salient object detection in images [49], and event detection from movies [48] and Twitter [17]. Furthermore, an approximator of the divergence between the joint distribution and the product of marginal distributions can be used for solving a wide range

of machine learning problems [22], including independence testing [24], feature selection [34, 9], feature extraction [33, 41], canonical dependency analysis [12], object matching [44], independent component analysis [32], clustering [31, 15], and causality learning [43]. For this reason, accurately approximating a divergence between two probability distributions from their samples has been an important challenge in the statistics, information theory, and machine learning communities.

A naive way to approximate the divergence from P to P', denoted by $D(P\|P')$, is to first obtain estimators $\widehat{P}_{\mathcal{X}}$ and $\widehat{P}'_{\mathcal{X}'}$ of the distributions P and P' separately from their samples $\mathcal{X}$ and $\mathcal{X}'$, and then compute a plug-in approximator $D(\widehat{P}_{\mathcal{X}}\|\widehat{P}'_{\mathcal{X}'})$. However, this naive approach violates *Vapnik's principle* [39]:

> *If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.*

More specifically, if we know the distributions P and P', we can immediately know their divergence $D(P\|P')$. However, knowing the divergence $D(P\|P')$ does not necessarily imply knowing the distributions P and P', because different pairs of distributions can yield the same divergence values. Thus, estimating the distributions P and P' is more general than estimating the divergence $D(P\|P')$. Following Vapnik's principle, direct divergence approximators $\widehat{D}(\mathcal{X}, \mathcal{X}')$ that do not involve the estimation of distributions P and P' have been developed recently [29, 18, 10, 47, 28].

The purpose of this particle is to give an overview of such direct divergence approximators.

2 Divergence Measures

In this section, we introduce useful divergence measures.

Kullback-Leibler (KL) Divergence: The most popular divergence measure in statistics and machine learning would be the KL divergence [16] defined as

$$\mathrm{KL}(p\|p') := \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} \mathrm{d}\boldsymbol{x},$$

where $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ are probability density functions of P and P', respectively.

Advantages of the KL divergence are that it is compatible with maximum likelihood estimation, it is invariant under input metric change, its Riemannian geometric structure is well studied [2], and it can be approximated accurately via *direct density-ratio estimation* [29, 18, 26]. However, it is not symmetric, it does not satisfy the triangle inequality, its approximation is computationally expensive due to the log function, and it is sensitive

to outliers and numerically unstable because of the strong non-linearity of the log function and possible unboundedness of the density-ratio function p/p' [4, 47].

Pearson (PE) Divergence: The PE divergence [19] is a squared-loss variant of the KL divergence defined as

$$\mathrm{PE}(p\|p') := \int p'(\boldsymbol{x}) \left(\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} - 1 \right)^2 \mathrm{d}\boldsymbol{x}.$$

Because both the PE and KL divergences belong to the class of Ali-Silvey-Csiszár divergences (which is also known as f-divergences) [1, 6], they share similar theoretical properties such as invariance under input metric change. The quadratic function the PE divergence adopts is compatible with least-squares estimation.

The PE divergence can also be accurately approximated via direct density-ratio estimation in the same way as the KL divergence [10, 26], but its approximator can be obtained *analytically* in a computationally much more efficient manner than the KL divergence. Furthermore, the PE divergence tends to be more robust against outliers than the KL divergence [27]. However, other weaknesses of the KL divergence such as asymmetry, violation of the triangle inequality, and possible unboundedness of the density-ratio function p/p' remain unsolved in the PE divergence.

Relative Pearson (rPE) Divergence: To overcome the possible unboundedness of the density-ratio function p/p', the rPE divergence was introduced recently [47], which is defined as

$$\mathrm{rPE}(p\|p') := \mathrm{PE}(p\|q_\alpha) = \int q_\alpha(\boldsymbol{x}) \left(\frac{p(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})} - 1 \right)^2 \mathrm{d}\boldsymbol{x},$$

where $q_\alpha = \alpha p + (1-\alpha)p'$ for $0 \leq \alpha < 1$. When $\alpha = 0$, the rPE divergence is reduced to the plain PE divergence. The quantity p/q_α is called the *relative density ratio*, which is always upper-bounded by $1/\alpha$ for $\alpha > 0$. Thus, it can overcome the unboundedness problem of the PE divergence, while the invariance under input metric change is still maintained.

The rPE divergence is still compatible with least-squares estimation, and it can be approximated in almost the same way as the PE divergence via *direct relative density-ratio estimation*. Indeed, an rPE divergence approximator can still be obtained analytically in an accurate and computationally efficient manner. However, it still violates symmetry and the triangle inequality in the same way as the KL and PE divergence, and the choice of α is not straightforward in practice.

L^2-Distance: The L^2-distance is another standard distance measure between probability distributions defined as

$$L^2(p, p') := \int \Big(p(\boldsymbol{x}) - p'(\boldsymbol{x}) \Big)^2 \mathrm{d}\boldsymbol{x}.$$

The L^2-distance is a proper distance measure, and thus it is symmetric and satisfies the triangle inequality. Furthermore, the density difference $p(\boldsymbol{x}) - p'(\boldsymbol{x})$ is always bounded as

long as each density is bounded. Therefore, the L^2-distance is stable, without the need of tuning any control parameter such as α in the rPE divergence.

The L^2-distance is also compatible with least-squares estimation, and it can be accurately and analytically approximated in a computationally efficient and numerically stable manner via *direct density-difference estimation* [28]. However, the L^2-distance is not invariant under input metric change, which is a unique property inherent to ratio-based divergences.

3 Direct Divergence Approximation

In this section, we review recent advances in direct divergence approximation.

KL Divergence Approximation [29]: The key idea is to estimate the density ratio p/p' without estimating the densities p and p'. More specifically, a density ratio approximator $\widehat{r}$ is obtained by minimizing the empirical KL divergence from p to $r \cdot p'$ with respect to a density-ratio model r:

$$\widehat{r} := \operatorname*{argmin}_{r} \frac{1}{n}\sum_{i=1}^{n} \log r(\boldsymbol{x}_i) \quad \text{subject to } r \geq 0 \text{ and } \frac{1}{n'}\sum_{i'=1}^{n'} r(\boldsymbol{x}'_{i'}) = 1.$$

For a linear-in-parameter density-ratio model defined by

$$r(\boldsymbol{x}) = \sum_{i=1}^{n} \theta_i \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2\sigma^2}\right), \tag{1}$$

the above optimization problem is convex and thus the global optimal solution can be obtained easily, e.g., by a gradient-projection iteration. The Gaussian width σ can be tuned by cross-validation with respect to the objective function. Given the density ratio estimator $\widehat{r}$, a KL divergence estimator $\widehat{\mathrm{KL}}(\mathcal{X}\|\mathcal{X}')$ can be constructed as

$$\widehat{\mathrm{KL}}(\mathcal{X}\|\mathcal{X}') := \frac{1}{n}\sum_{i=1}^{n} \log \widehat{r}(\boldsymbol{x}_i).$$

A MATLAB® implementation of the above KL divergence approximator (called the *KL importance estimation procedure*; KLIEP) is available from

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

Variations of this procedure for various density ratio models have been developed, including the log-linear model [38], the Gaussian mixture model [42], and the mixture of probabilistic principal component analyzers [46]. Also, an unconstrained variant, which corresponds to approximately maximizing the *Legendre-Fenchel lower bound* of the KL divergence [14], was also proposed [18]:

$$\widehat{\mathrm{KL}}'(\mathcal{X}\|\mathcal{X}') := \max_{r} \left[\frac{1}{n}\sum_{i=1}^{n} \log r(\boldsymbol{x}_i) - \frac{1}{n'}\sum_{i'=1}^{n'} r(\boldsymbol{x}'_{i'}) + 1\right].$$

PE Divergence Approximation [10]: The PE divergence can also be directly approximated without estimating the densities p and p' via direct estimation of the density ratio p/p'. More specifically, a density ratio approximator $\widehat{r}$ is obtained by minimizing the empirical p'-weighted squared difference between a density ratio model r and the true density ratio p/p':

$$\widehat{r} := \operatorname*{argmin}_{r} \left[\frac{1}{n'} \sum_{i'=1}^{n'} r^2(\boldsymbol{x}'_{i'}) - \frac{2}{n} \sum_{i=1}^{n} r(\boldsymbol{x}_i) \right].$$

For the linear-in-parameter density-ratio model (1) possibly together with the ℓ_2-regularization [8], the density ratio estimator $\widehat{r}$ can be obtained analytically, with a closed-form leave-one-out cross-validation score [40]. Moreover, together with the ℓ_1-regularization [35], the coefficients $\{\theta_i\}_{i=1}^n$ tend to be sparse and can be learned in a computationally efficient way [36], further equipped with a regularization path tracking algorithm [7].

A MATLAB® implementation with the ℓ_2-regularizer (called *unconstrained least-squares importance fitting*; uLSIF) is available from

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

rPE Divergence Approximation [47]: The rPE divergence can be estimated in the same way as the PE divergence as

$$\widehat{r} := \operatorname*{argmin}_{r} \left[\frac{\alpha}{n} \sum_{i=1}^{n} r^2(\boldsymbol{x}_i) + \frac{1-\alpha}{n'} \sum_{i'=1}^{n'} r^2(\boldsymbol{x}'_{i'}) - \frac{2}{n} \sum_{i=1}^{n} r(\boldsymbol{x}_i) \right].$$

Thus, all the computational advantages of PE divergence approximation mentioned above are inherited to rPE divergence approximation.

A MATLAB® implementation of this algorithm (called *relative uLSIF*; RuLSIF) is available from

"`http://sugiyama-www.cs.titech.ac.jp/~yamada/RuLSIF.html`".

L^2-Distance Approximation [28]: The key idea is to directly estimate the density difference $p - p'$ without estimating each density. More specifically, a density difference approximator $\widehat{f}$ is obtained by minimizing the empirical squared difference between a density difference model f and the true density difference $p - p'$:

$$\widehat{f} := \operatorname*{argmin}_{f} \left[\int f(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} - \left(\frac{2}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i) - \frac{2}{n'} \sum_{i'=1}^{n'} f(\boldsymbol{x}'_{i'}) \right) \right].$$

In practice, the use of the Gaussian kernel model,

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \theta_i \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2\sigma^2}\right) + \sum_{i'=1}^{n'} \theta_{n+i'} \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'_{i'}\|^2}{2\sigma^2}\right),$$

is advantageous because the first term $\int f(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x}$ in the objective function can be computed analytically for this model. The above optimization problem is essentially the same form as least-squares density-ratio approximation for the PE divergence, and therefore least-squares density-difference approximation can enjoy all the computational properties of least-squares density-ratio approximation.

A MATLAB® implementation of the above algorithm (called *least-squares density difference*; LSDD) is available from

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

Convergence Issues: All the direct divergence approximators reviewed above were proved to achieve the $\sqrt{n}$-consistency in the parametric case (suppose $n' = n$) [29, 10, 47, 28], which is the optimal convergence rate. Furthermore, they were also proved to achieve the mini-max optimal convergence rate in the non-parametric case [18, 29, 10, 47, 28]. Also experimentally, direct divergence approximators were shown to outperform the naive approaches based on density estimation [29, 10, 47, 28].

4 Usage of Divergence Estimators in Machine Learning

In this section, we show applications of divergence estimators in machine learning.

Change-Detection in Time-Series: The goal is to discover abrupt property changes behind time-series data. Let $\boldsymbol{y}(t) \in \mathbb{R}^m$ be an m-dimensional time-series sample at time t, and let $\boldsymbol{Y}(t) := [\boldsymbol{y}(t)^\top, \boldsymbol{y}(t+1)^\top, \ldots, \boldsymbol{y}(t+k-1)^\top]^\top \in \mathbb{R}^{km}$ be a subsequence of time series at time t with length k. Instead of a single point $\boldsymbol{y}(t)$, the subsequence $\boldsymbol{Y}(t)$ is treated as a sample here, because time-dependent information can be naturally incorporated by this trick [13]. Let $\mathcal{Y}(t) := \{\boldsymbol{Y}(t), \boldsymbol{Y}(t+1), \ldots, \boldsymbol{Y}(t+r-1)\}$ be a set of r retrospective subsequence samples starting at time t. Then a divergence between the probability distributions of $\mathcal{Y}(t)$ and $\mathcal{Y}(t+r)$ may be used as the plausibility of change points (see Figure 1).

The change-detection methods based on the rPE divergence [17] and the L^2-distance [28] were shown to be promising through experiments. In particular, the method based on the rPE divergence was successfully applied to event detection from movies [48] and Twitter [17].

Class-Prior Estimation under Class-Balance Change: In real-world pattern recognition tasks, changes in class balance are often observed between the training and test

Figure 1: Change-point detection in time-series.

Figure 2: Class-prior estimation.

phases. In such cases, naive classifier training produces significant estimation bias because the class balance in the training dataset does not properly reflect that of the test dataset. Here, let us consider a binary pattern recognition task of classifying pattern $\boldsymbol{x} \in \mathbb{R}^d$ to class $y \in \{+1, -1\}$. The goal is to learn the class balance of a test dataset in a semi-supervised learning setup where unlabeled test samples are provided in addition to labeled training samples [3]. The class balance in the test set can be estimated by matching a mixture of class-wise training input densities,

$$q_{\mathrm{test}}(\boldsymbol{x}) := \pi p_{\mathrm{train}}(\boldsymbol{x}|y = +1) + (1 - \pi) p_{\mathrm{train}}(\boldsymbol{x}|y = -1),$$

to the test input density $p_{\mathrm{test}}(\boldsymbol{x})$ under some divergence measure [20]. Here, $\pi \in [0, 1]$ is a mixing coefficient to be learned to minimize the divergence (see Figure 2).

The class-balance estimation methods based on the PE divergence [20] and the L^2-distance [28] were shown to be promising through experiments.

Salient Object Detection in an Image: The goal is to find salient objects in an image. This can be achieved by computing a divergence between the probability distributions of image features (such as brightness, edges, and colors) in the center window and its surroundings [49]. This divergence computation is swept over the entire image, possibly with changing scales (Figure 3).

Figure 3: Object detection in an image.

The object detection method based on the rPE divergence was demonstrated to be promising in experiments [49].

Measuring Statistical Independence: The goal is to measure how strongly two random variables $\boldsymbol{U}$ and $\boldsymbol{V}$ are statistically dependent, using paired samples $\{(\boldsymbol{u}_i, \boldsymbol{v}_i)\}_{i=1}^n$ drawn independently from the joint distribution with density $p_{\mathbf{U},\mathbf{V}}(\boldsymbol{u}, \boldsymbol{v})$. Let us consider a divergence between the joint density $p_{\mathbf{U},\mathbf{V}}$ and the product of marginal densities $p_{\mathbf{U}} \cdot p_{\mathbf{V}}$. This actually serves as a measure of statistical independence, because $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent if and only if the divergence is zero (i.e., $p_{\mathbf{U},\mathbf{V}} = p_{\mathbf{U}} \cdot p_{\mathbf{V}}$), and the dependence between $\boldsymbol{U}$ and $\boldsymbol{V}$ is stronger if the divergence is larger.

Such a dependence measure can be approximated in the same way as ordinary divergences by using the two datasets formed as $\mathcal{X} = \{(\boldsymbol{u}_i, \boldsymbol{v}_i)\}_{i=1}^n$ and $\mathcal{X}' = \{(\boldsymbol{u}_i, \boldsymbol{v}_j)\}_{i,j=1}^n$. The dependence measure based on the KL divergence is called *mutual information* [21], which plays a central role in information theory [5]. On the other hand, its PE divergence variant is called the *squared-loss mutual information*, which was shown to be useful for solving various machine learning tasks [22] such as independence testing [24], feature selection [34, 9], feature extraction [33, 41], canonical dependency analysis [12], object matching [44], independent component analysis [32], clustering [31, 15], and causality learning [43]. An L^2-distance variant of the dependence measure is called *quadratic mutual information* [37].

5 Conclusions

In this article, we reviewed recent advances in direct divergence approximation. Direct divergence approximators theoretically achieve optimal convergence rates both in parametric and non-parametric cases and experimentally compare favorably with the naive density estimation counterparts. However, direct divergence approximators still suffer from the curse of dimensionality. A possible cure for this problem is to combine them width dimension reduction, based on the hope that two probability distributions share some commonality [23, 30, 45]. Further investigating this line would be a promising future

direction.

Acknowledgements: The author acknowledges support from the JST PRESTO program, KAKENHI 25700022, the FIRST program, and AOARD.

References

[1] Ali, S.M., Silvey, S.D.: A general class of coefficients of divergence of one distribution from another. Journal of the Royal Statistical Society, Series B **28**(1), 131–142 (1966)

[2] Amari, S., Nagaoka, H.: Methods of Information Geometry. Oxford University Press, Providence, RI, USA (2000)

[3] Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. MIT Press, Cambridge, MA, USA (2006)

[4] Cortes, C., Mansour, Y., Mohri, M.: Learning bounds for importance weighting. In: J. Lafferty, C.K.I. Williams, R. Zemel, J. Shawe-Taylor, A. Culotta (eds.) Advances in Neural Information Processing Systems 23, pp. 442–450 (2010)

[5] Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. John Wiley & Sons, Inc., Hoboken, NJ, USA (2006)

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

[7] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. The Annals of Statistics **32**(2), 407–499 (2004)

[8] Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(3), 55–67 (1970)

[9] Jitkrittum, W., Hachiya, H., Sugiyama, M.: Feature selection via ℓ_1-penalized squared-loss mutual information. IEICE Transactions on Information and Systems **E96-D**(7), 1513–1524 (2013)

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

[11] Kanamori, T., Suzuki, T., Sugiyama, M.: f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. IEEE Transactions on Information Theory **58**(2), 708–720 (2012)

[12] Karasuyama, M., Sugiyama: Canonical dependency analysis based on squared-loss mutual information. Neural Networks **34**, 46–55 (2012)

[13] Kawahara, Y., Sugiyama, M.: Sequential change-point detection based on direct density-ratio estimation. Statistical Analysis and Data Mining **5**(2), 114–127 (2012)

[14] Keziou, A.: Dual representation of ϕ-divergences and applications. Comptes Rendus Mathématique **336**(10), 857–862 (2003)

[15] Kimura, M., Sugiyama, M.: Dependence-maximization clustering with least-squares mutual information. Journal of Advanced Computational Intelligence and Intelligent Informatics **15**(7), 800–805 (2011)

[16] Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics **22**, 79–86 (1951)

[17] Liu, S., Yamada, M., Collier, N., Sugiyama, M.: Change-point detection in time-series data by relative density-ratio estimation. Neural Networks **43**, 72–83 (2013)

[18] Nguyen, X., Wainwright, M.J., Jordan, M.I.: Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Transactions on Information Theory **56**(11), 5847–5861 (2010)

[19] Pearson, K.: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine Series 5 **50**(302), 157–175 (1900)

[20] du Plessis, M.C., Sugiyama, M.: Semi-supervised learning of class balance under class-prior change by distribution matching. In: J. Langford, J. Pineau (eds.) Proceedings of 29th International Conference on Machine Learning (ICML2012), pp. 823–830. Edinburgh, Scotland (2012)

[21] Shannon, C.: A mathematical theory of communication. Bell Systems Technical Journal **27**, 379–423 (1948)

[22] Sugiyama, M.: Machine learning with squared-loss mutual information. Entropy **15**(1), 80–112 (2013)

[23] Sugiyama, M., Kawanabe, M., Chui, P.L.: Dimensionality reduction for density ratio estimation in high-dimensional spaces. Neural Networks **23**(1), 44–59 (2010)

[24] Sugiyama, M., Suzuki, T.: Least-squares independence test. IEICE Transactions on Information and Systems **E94-D**(6), 1333–1336 (2011)

[25] Sugiyama, M., Suzuki, T., Itoh, Y., Kanamori, T., Kimura, M.: Least-squares two-sample test. Neural Networks **24**(7), 735–751 (2011)

[26] Sugiyama, M., Suzuki, T., Kanamori, T.: Density Ratio Estimation in Machine Learning. Cambridge University Press, Cambridge, UK (2012)

[27] Sugiyama, M., Suzuki, T., Kanamori, T.: Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation. Annals of the Institute of Statistical Mathematics **64**(5), 1009–1044 (2012)

[28] Sugiyama, M., Suzuki, T., Kanamori, T., du Plessis, M.C., Liu, S., Takeuchi, I.: Density-difference estimation. Neural Computation **25**(10), 2734–2775 (2013)

[29] Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., Kawanabe, M.: Direct importance estimation for covariate shift adaptation. Annals of the Institute of Statistical Mathematics **60**(4), 699–746 (2008)

[30] Sugiyama, M., Yamada, M., von Bünau, P., Suzuki, T., Kanamori, T., Kawanabe, M.: Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. Neural Networks **24**(2), 183–198 (2011)

[31] Sugiyama, M., Yamada, M., Kimura, M., Hachiya, H.: On information-maximization clustering: Tuning parameter selection and analytic solution. In: L. Getoor, T. Scheffer (eds.) Proceedings of 28th International Conference on Machine Learning (ICML2011), pp. 65–72. Bellevue, Washington, USA (2011)

[32] Suzuki, T., Sugiyama, M.: Least-squares independent component analysis. Neural Computation **23**(1), 284–301 (2011)

[33] Suzuki, T., Sugiyama, M.: Sufficient dimension reduction via squared-loss mutual information estimation. Neural Computation **3**(25), 725–758 (2013)

[34] Suzuki, T., Sugiyama, M., Kanamori, T., Sese, J.: Mutual information estimation reveals global associations between stimuli and biological processes. BMC Bioinformatics **10**(1), S52 (12 pages) (2009)

[35] Tibshirani, R.: Regression shrinkage and subset selection with the lasso. Journal of the Royal Statistical Society, Series B **58**(1), 267–288 (1996)

[36] Tomioka, R., Suzuki, T., Sugiyama, M.: Super-linear convergence of dual augmented Lagrangian algorithm for sparsity regularized estimation. Journal of Machine Learning Research **12**, 1537–1586 (2011)

[37] Torkkola, K.: Feature extraction by non-parametric mutual information maximization. Journal of Machine Learning Research **3**, 1415–1438 (2003)

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

[39] Vapnik, V.N.: Statistical Learning Theory. Wiley, New York, NY, USA (1998)

[40] Wahba, G.: Spline Models for Observational Data. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (1990)

[41] Yamada, M., Niu, G., Takagi, J., Sugiyama, M.: Computationally efficient sufficient dimension reduction via squared-loss mutual information. In: C.N. Hsu, W.S. Lee (eds.) Proceedings of the Third Asian Conference on Machine Learning (ACML2011), *JMLR Workshop and Conference Proceedings*, vol. 20, pp. 247–262. Taoyuan, Taiwan (2011)

[42] Yamada, M., Sugiyama, M.: Direct importance estimation with Gaussian mixture models. IEICE Transactions on Information and Systems **E92-D**(10), 2159–2162 (2009)

[43] Yamada, M., Sugiyama, M.: Dependence minimizing regression with model selection for non-linear causal inference under non-Gaussian noise. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI2010), pp. 643–648. The AAAI Press, Atlanta, Georgia, USA (2010)

[44] Yamada, M., Sugiyama, M.: Cross-domain object matching with model selection. In: G. Gordon, D. Dunson, M. Dudík (eds.) Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS2011), *JMLR Workshop and Conference Proceedings*, vol. 15, pp. 807–815. Fort Lauderdale, Florida, USA (2011)

[45] Yamada, M., Sugiyama, M.: Direct density-ratio estimation with dimensionality reduction via hetero-distributional subspace analysis. In: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI2011), pp. 549–554. The AAAI Press, San Francisco, California, USA (2011)

[46] Yamada, M., Sugiyama, M., Wichern, G., Simm, J.: Direct importance estimation with a mixture of probabilistic principal component analyzers. IEICE Transactions on Information and Systems **E93-D**(10), 2846–2849 (2010)

[47] Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., Sugiyama, M.: Relative density-ratio estimation for robust distribution comparison. Neural Computation **25**(5), 1324–1370 (2013)

[48] Yamanaka, M., Matsugu, M., Sugiyama, M.: Detection of activities and events without explicit categorization. IPSJ Transactions on Mathematical Modeling and Its Applications **6**(2), 86–92 (2013)

[49] Yamanaka, M., Matsugu, M., Sugiyama, M.: Salient object detection based on direct density-ratio estimation. IPSJ Transactions on Mathematical Modeling and Its Applications **6**(2), 78–85 (2013)