ESTIMATORS FOR UNNORMALIZED STATISTICAL MODELS BASED ON SELF DENSITY RATIO

Kazuyuki Hiraoka† *Toshihiko Hamada*† *Gen Hori*‡§

†Wakayama National College of Technology, Wakayama 644-0023, Japan
‡Faculty of Business Administration, Asia University, Tokyo 180-8629, Japan
§Brain Science Institute, RIKEN, Saitama 351-0198, Japan
E-mail : hiraoka@wakayama-nct.ac.jp

ABSTRACT

A wide family of consistent estimators is introduced for unnormalized statistical models. They do not need normalization of the probability density function (PDF) because they are based on the density ratio between the same PDF at different points; the multiplicative normalization constant is canceled there. We construct a family of estimators based on pairwise comparison of density ratio and derive several estimators as its special cases. The family includes score matching as its parameter limit and outperforms score matching for the optimal value of the parameter. We share the idea of random transformations with contrastive divergence whereas we do not assume Markov chain and obtain consistent deterministic estimators by analytic averaging.

Index Terms— unnormalized statistical models, self density ratio, score matching, consistency

1. INTRODUCTION

Intractability of normalization is a major computational bottleneck of machine learning with complex statistical models nowadays. The probability distribution is known only up to a multiplicative normalization constant in that case. A typical example is deep learning where contrastive divergence (CD) [16] is used to circumvent this difficulty [25]. Similar problems also appear in various applications [12] including graphical models [15], unsupervised feature learning [2], computational neuroscience [13] and modeling of images [20].

Approaches to estimation of unnormalized statistical models are classified as follows. Typical strategies to avoid normalization are elimination, estimation and approximation. The normalization constant is eliminated by exploiting criteria based on $(\log f)'$ or $f(v)/f(u)$ for a probability density function (PDF) $f(x)$ in the first strategy whereas it is estimated as an unknown parameter in the second strategy. In the third strategy, some kind of approximation is used instead of $f(x)$ itself. The approaches are also divided into deterministic methods and Monte Carlo methods. Pseudorandom numbers are used in the latter methods and the result changes for every run. Table 1 classifies conventional methods such as score matching (SM) [17], generalized score matching (GSM) [22], ratio matching (RM) [18], local scoring rules (LSR) [21], noise-contrastive estimation (NCE) [11], Monte Carlo maximum likelihood (MCML) estimation [9], contrastive divergence (CD) [16], minimum probability flow learning (MPF) [26], pseudolikelihood (PL) [3], Mean field (MF) theory and variational Bayes (VB) techniques [1].

Table 1. Approaches to estimation of unnormalized models

	Elimination	Estimation	Approximation
Deterministic	(G)SM, RM, LSR		MF, VB, PL, MPF
Monte Carlo		NCE	MCML, CD

The goal of this paper is to construct an estimation method that satisfies the following requests:

(a) It does not need normalization of the density function.
(b) It is consistent; the estimation converges to the true value if the sample size increases indefinitely.
(c) It is deterministic for convenience; Monte Carlo simulation is not required.
(d) It yields a wide family of estimators so that we can find a good one for each problem.
(e) It has intuitive interpretations that help understanding of its behavior.

Taking deterministic elimination approach in Table 1, we give a family of consistent estimators of such models for continuous data. Our key idea is extensive use of the density ratio between the same PDF $f(x)$ at different points. The density ratio is attracting a great deal of attention [8, 27, 28] in various statistical data processing tasks.

The rest of the paper is organized as follows. Section 2 introduces the unnormalized statistical model and presents the shift-based and scaling-based estimators. Section 3 constructs estimators based on pairwise comparison. Section 4 presents generic form of estimators from which all the proposed methods in this paper are derived. Section 5 gives an inference example to demonstrate that our proposed method outperforms score matching. Section 6 contains concluding remarks.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. ESTIMATION VIA SHIFT AND SCALING

2.1. Unnormalized statistical model

We consider estimation of a unnormalized statistical model, i.e. a family of PDF p on a random variable X,

$$f(x;\theta) \equiv p(X = x;\theta) = \frac{1}{C(\theta)} F(x;\theta), \qquad x \in \mathcal{X}$$

with a given function $F(x;\theta) \geq 0$ and the unknown parameter $\theta \in \Theta$. The calculation of

$$C(\theta) \equiv \int_{\mathcal{X}} F(x;\theta)\, dx$$

is assumed to be intractable. Let $\theta^* \in \Theta$ be the true parameter and we desire to estimate it from i.i.d. samples $x_1, \ldots, x_n \sim f(\,\cdot\,;\theta^*)$ without using $C(\theta)$ explicitly. For simplicity, we suppose that the data space $\mathcal{X}$ and the parameter space Θ are subsets of $\mathbb{R}$, the set of real numbers.

2.2. Penalty criterion for estimation

We consider estimators $\hat{\theta}$ of the form

$$\hat{\theta} \equiv \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_{t} M(x_t;\theta) \tag{1}$$

for some penalty criterion $M(x;\theta)$, that is called a scoring rule [21]. Our goal is to construct a good criterion $M(x;\theta)$ that produces a good estimator. We require the consistency of the estimator for which it is necessary that the true parameter $\theta = \theta^*$ is a critical point of the expectation $E[M(X;\theta)]$,

$$E\left[\left. \frac{\partial M(X;\theta)}{\partial \theta} \right|_{\theta=\theta^*} \right] = \int_{\mathcal{X}} f(x;\theta^*) \left. \frac{\partial M(x;\theta)}{\partial \theta} \right|_{\theta=\theta^*} dx = 0$$

for $X \sim f(\,\cdot\,;\theta^*)$. This unbiasedness implies the consistency as long as usual regularity conditions are satisfied [19].

2.3. Estimation via self density ratio

An invertible mapping $\phi : \mathcal{X} \to \mathcal{X}$ on the data space plays a key role in our approach. When we observe a sample x, we use the self density ratio

$$f(x;\theta) : f(\phi(x);\theta) = F(x;\theta) : F(\phi(x);\theta)$$

for the estimation. Note that we can calculate it without the normalization constant $C(\theta)$. For notational convenience, we construct various penalty criteria $M(x;\theta)$ based on the following two quantities,

$$\xi(x;\theta) \equiv \frac{\phi'(x) F(\phi(x);\theta)}{F(x;\theta)}, \quad \eta(x;\theta) \equiv \frac{\psi'(x) F(\psi(x);\theta)}{F(x;\theta)} \tag{2}$$

where $\psi \equiv \phi^{-1}$ is the inverse of the mapping ϕ. The derivatives $\phi'(x)$ and $\psi'(x)$ appear naturally in the transformation of PDF. We also suppose $\phi'(x) > 0$ for simplicity.

2.4. Shift-based and scaling-based estimators

A typical penalty criterion based on self density ratio is of the form,

$$M(x;\theta) \equiv \log\bigl(1 + \xi(x;\theta)\bigr) + \log\bigl(1 + \eta(x;\theta)\bigr) \tag{3}$$

which is a special case of the penalty criterion (8) constructed based on pairwise comparison in Section 3.2. We obtain (3) by setting a uniform weight $\rho(u, u', \omega) = 1$, $\omega = 1$, $u' = \phi(u)$ and $L_1(q) = L_0(q) = -\log q$ in (8).

In particular, we obtain the shift-based criterion

$$M_a^{\text{shift}}(x) \equiv \log\left(1 + \frac{f(x+a)}{f(x)}\right) + \log\left(1 + \frac{f(x-a)}{f(x)}\right) \tag{4}$$

by setting $\phi(x) = x + a$ for a positive constant $a > 0$ in (3). The estimator based on the criterion (4) coincides with score matching [17] for the limit $a \to 0$. Also we obtain the scaling-based criterion

$$M_a^{\text{scaling}}(x) \equiv \log\left(1 + \frac{a f(ax)}{f(x)}\right) + \log\left(1 + \frac{a^{-1} f(a^{-1}x)}{f(x)}\right) \tag{5}$$

by setting $\phi(x) = ax$ for a constant $a > 1$ in (3). The latter is suitable for a positive data space $\mathcal{X} = (0, \infty)$. We omit θ in (4) and (5) for simplicity. Section 5 gives inference example using (4) to show that the estimator based on the criterion outperforms score matching [17] for the optimal value of the parameter a. Section 5 also introduces leave-worst-one-out method to find the optimal parameter a.

3. ESTIMATOR CONSTRUCTION VIA PAIRWISE COMPARISON

3.1. Pairwise comparison of density ratio

We start from naive estimations based on pairwise comparisons of densities and then construct a new estimator as a weighted sum of them in the following section.

Consider two disjoint infinitesimal intervals on $\mathbb{R}$,

$$I \equiv [u, u + \Delta u], \quad I' \equiv [u', u' + \Delta u'],$$

and let n_I and $n_{I'}$ be the numbers of observed samples in I and I' respectively. The ratio $n_I : n_{I'}$ gives a trivial estimation of the probability ratio $P(X \in I;\theta) : P(X \in I';\theta)$. We then make an estimate $\theta = \hat{\theta}_{I,I'}$ so that these ratios are as close as possible. Of course this is extremely inefficient because most samples outside the intervals are wasted. Yet it should work theoretically if $n \to \infty$ in the sense that the estimate converges to $\tilde{\theta}$ such that $P(X \in I;\tilde{\theta}) : P(X \in I';\tilde{\theta}) = P(X \in I;\theta^*) : P(X \in I';\theta^*)$.

This estimation is formulated as maximum likelihood estimation (MLE) as follows. Suppose

$$\begin{aligned} Q(I, I'; \theta) &\equiv P(X \in I \,|\, X \in I \cup I'; \theta) \\ &= \frac{P(X \in I; \theta)}{P(X \in I; \theta) + P(X \in I'; \theta)} \\ &\approx \frac{f(u;\theta)\Delta u}{f(u;\theta)\Delta u + f(u';\theta)\Delta u'}, \end{aligned}$$

then the estimate is expressed as

$$\hat{\theta}_{I,I'} \equiv \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_t l_{I,I'}(x_t; \theta), \tag{6}$$

$$l_{I,I'}(x;\theta) \equiv \begin{cases} L_1\big(Q(I, I'; \theta)\big) & (x \in I) \\ L_0\big(Q(I', I; \theta)\big) & (x \in I') \\ 0 & (\text{otherwise}) \end{cases}$$

where L_1 and L_0 forms a pair of loss functions that satisfies the properness

$$\operatorname*{argmin}_{q} \{pL_1(q) + (1-p)L_0(1-q)\} = p \tag{7}$$

for $0 < p < 1$ and $0 < q < 1$. By the law of large numbers, the sample mean of $l_{I,I'}$ in (6) converges to its expectation

$$\begin{aligned} E[l_{I,I'}(X;\theta)] &= P(X \in I; \theta^*) L_1\big(Q(I, I'; \theta)\big) \\ &\quad + P(X \in I'; \theta^*) L_0\big(Q(I', I; \theta)\big) \\ &\propto Q(I, I'; \theta^*) L_1\big(Q(I, I'; \theta)\big) \\ &\quad + Q(I', I; \theta^*) L_0\big(Q(I', I; \theta)\big) \end{aligned}$$

for $n \to \infty$. The right hand side is minimized at $\theta = \theta^*$ because of the properness (7). In particular, we obtain MLE if we use the logarithmic score for the loss function,

$$L_1(q) = L_0(q) = -\log q$$

that satisfies (7). Another example of the loss function is the Brier score [5], $L_1(q) = L_0(q) = (1-q)^2$.

3.2. Estimator based on pairwise comparison

Based on the naive estimations with pairwise comparisons, we construct a estimator by summing them with changing the intervals.

Considering the limit of $l_{I,I'}(x;\theta)/\Delta u$ with $\Delta u \to +0$ and $\omega \equiv \Delta u'/\Delta u > 0$, we obtain

$$\begin{aligned} l_{u,u',\omega}(x;\theta) &\equiv \delta(x-u) L_1\big(Q(u, u', \omega; \theta)\big) \\ &\quad + \omega\delta(x-u') L_0\big(Q(u', u, \omega^{-1}; \theta)\big) \end{aligned}$$

with

$$\begin{aligned} Q(u, u', \omega; \theta) &\equiv \frac{f(u;\theta)}{f(u;\theta) + \omega f(u';\theta)} \\ &= \frac{F(u;\theta)}{F(u;\theta) + \omega F(u';\theta)} \end{aligned}$$

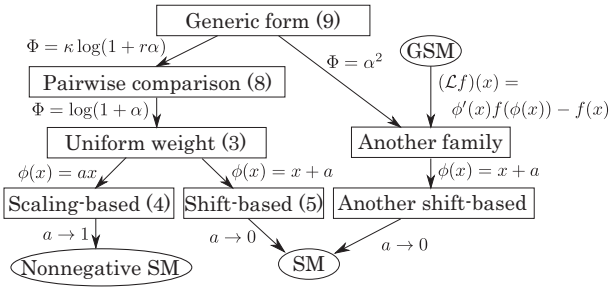

Fig. 1. The relations among the proposed methods (shown in boxes) and previous methods (shown in ovals). The arrow $A \to B$ indicates that B is a special case of A. "Another family" and "Another shift-based" are not mentioned in the paper and will be presented in our forthcoming paper.

where $\delta(x)$ is the Dirac delta function. Note that we can calculate $Q(u, u', \omega; \theta)$ without the normalization constant $C(\theta)$. Now we construct a penalty criterion $M(x;\theta)$ based on a weighted sum of $l_{u,u',\omega}$ with an arbitrary weight function $\rho(u, u', \omega) \geq 0$,

$$\begin{aligned} &M(x;\theta) \\ &\equiv \int_0^\infty \int_{\mathcal{X}} \int_{\mathcal{X}} \rho(u, u', \omega) l_{u,u',\omega}(x;\theta)\, du\, du'\, d\omega \\ &= \int_0^\infty \int_{\mathcal{X}} \rho(x, u', \omega) L_1\big(Q(x, u', \omega; \theta)\big)\, du'\, d\omega \\ &\quad + \int_0^\infty \int_{\mathcal{X}} \omega\rho(u, x, \omega) L_0\big(Q(x, u, \omega^{-1}; \theta)\big)\, du\, d\omega. \end{aligned} \tag{8}$$

The consistency follows from the consistency of MLE [19].

In particular, we obtain the penalty criterion (3) presented in Section 2.4 by setting $\rho(u, u', \omega) = 1$, $\omega = 1$, $u' = \phi(u)$ and $L_1(q) = L_0(q) = -\log q$ in (8).

4. GENERIC FORM OF ESTIMATORS

In this section, we present a generic form of penalty criterion from which all the proposed methods in this paper are derived as special cases. Fig. 1 illustrates the relations among the proposed methods and some previous methods in Table 1. The related proofs are omitted due to space limitations and will be presented in our forthcoming paper.

The generic form of penalty criterion is given based on ξ and η defined in Section 2.3 as

$$M(x;\theta) = \Phi(\xi(x;\theta), x) + \Psi(\eta(x;\theta), x) \tag{9}$$

for an arbitrary pair of functions Φ and Ψ that satisfy

$$\Phi'(\alpha, x) \equiv \frac{\partial \Phi(\alpha, x)}{\partial \alpha} = \frac{1}{\alpha} \Psi'\left(\frac{1}{\alpha}, \phi(x)\right), \tag{10}$$

$$\Psi'(\beta, x) \equiv \frac{\partial \Psi(\beta, x)}{\partial \beta} = \frac{1}{\beta} \Phi'\left(\frac{1}{\beta}, \psi(x)\right). \tag{11}$$

where x, $\phi(x)$, $\psi(x)$, $\phi(\phi(x))$ and $\psi(\psi(x))$ are different from one another. The conditions (10) and (11) are equivalent because the latter is obtained from the former by replacing α with $1/\beta$ and x with $\psi(x)$. For example, the penalty criterion (3) presented in Section 2.4 is obtained by setting $\Phi(\alpha, x) = \log(1+\alpha)$ and $\Psi(\beta, x) = \log(1+\beta)$ in (9).

5. INFERENCE EXAMPLE

We present an inference example with the shift-based criterion (4) to demonstrate that our proposed method outperforms score matching [17] for the optimal value of the parameter a. The shift-based estimator coincides with score matching for the limit $a \to 0$ but the optimal value of a is not 0. The optimal value of a is selected automatically by the leave-worst-one-out method.

The benchmark is a mixture of two Gaussian distributions $N(0,1)$ and $N(6,1)$ with unknown weights $(1-\theta)$ and θ,

$$f(x;\theta) = (1-\theta)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} + \theta\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-6)^2}$$

where $0 \le \theta \le 1$. The estimation errors are shown in Fig. 2 for MLE, score matching and the proposed shift-based estimator (4). The optimal $\hat{\theta}_a$ for the shift $a = 6$ is almost equivalent to MLE and better than score matching. We further try automatic selection of a and its result $\hat{\theta}_{\hat{a}}$ is also better than score matching.

We carry out the leave-worst-one-out method for automatic selection of a as follows. Let the worst sample $x_{\text{worst}(a)}$ for the given a be the sample x_t that gives the largest $|\partial M_a^{\text{shift}}(x_t;\theta)/\partial\theta|$ at $\theta = \hat{\theta}_a$. Since $\hat{\theta}$ is determined from $\sum_t \partial M_a^{\text{shift}}(x_t;\theta)/\partial\theta = 0$, $x_{\text{worst}(a)}$ is expected to give the largest influence to estimation. Then we repeat estimation once again from the samples $x_1, \ldots, x_n$ except for $x_{\text{worst}(a)}$ and obtain $\hat{\theta}_a^{\text{worst}}$. We finally select $a = \hat{a}$ such that $|\hat{\theta}_a^{\text{worst}} - \hat{\theta}_a|$ is minimized.

In our simulations, linear approximation and the central limit theorem are not applicable in detection of extremely bad a. The criterion M_a^{shift} is strongly nonlinear and outliers give large contribution in minimization for such a.

6. CONCLUDING REMARKS

We have introduced a family of consistent estimators of unnormalized statistical models that includes score matching as its limit and shown that the optimal one in the family outperforms score matching using an inference example. The key of our approach is extensive use of the density ratio between the same PDF $f(x)$ at different points. A transformation on the data space plays an important role; it controls selection of the points for the density ratio. Our future study includes strategies for the selection of a good transformation ϕ as well as the selection of the generator Φ in Section 4.

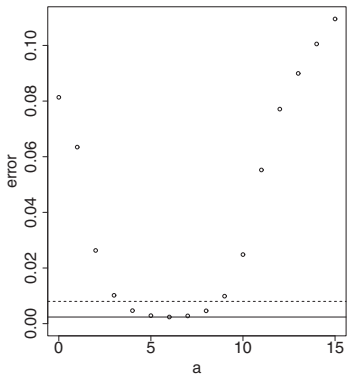

Fig. 2. The mean of the squared error $(\hat{\theta} - \theta^*)^2$ of 1000 sessions for the sample size $n = 100$ and the true parameter $\theta^* = 0.3$ in the Gaussian mixture model. The leftmost point at $a = 0$ corresponds to score matching and other points the proposed shift-based estimator (4) for the shift $a = 1, 2, \ldots, 15$. The solid and dashed lines indicate the errors of $\hat{\theta}_{\text{MLE}}$ and $\hat{\theta}_{\hat{a}}$ with automatically selected $\hat{a}$ based on the leave-worst-one-out method respectively. A numerical minimization is used for each estimation.

Note that the shift-based criterion (4) can be applied to multidimensional data $x \in \mathbb{R}^d$. We may use the estimator

$$\hat{\theta} \equiv \operatorname*{argmin}_{\theta} \sum_{i=1}^{n}\sum_{j=1}^{k} M_{a_j}^{\text{shift}}(x_i)$$

with known constant vectors $a_1, \ldots, a_k \in \mathbb{R}^d$ in that case. Our proposed methods can be also applied to the case of discrete $x \in \mathcal{D}$ in the same way with a transformation $\phi : \mathcal{D} \to \mathcal{D}$ on the data space $\mathcal{D}$. It yields estimators based on the logarithm of probabilities that seem different from generalized score matching [22] and ratio matching [18] based on square distances.

7. REFERENCES

[1] H. Attias, "A variational Bayesian framework for graphical models," *Advances in Neural Information Processing Systems*, vol. 12, pp. 209–215, 2000.

[2] Y. Bengio, A. C. Courville, and P. Vincent, "Unsupervised feature learning and deep learning: a review and new perspectives," *arXiv*, vol. 1206.5538 [cs.LG], 2012.

[3] J. Besag, "Statistical analysis of non-lattice data," *The Statistician*, vol. 24, pp. 179–195, 1975.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[5] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Rev.*, vol. 78, pp. 1–3, 1950.

[6] M. Á. Carreira-Perpigñán and G. E. Hinton, "On contrastive divergence learning," *Artificial Intelligence and Statistics*, 2005.

[7] R. O. Duda, P. E. Hart, and D. G. Stock, *Pattern Classification*, 2nd ed., John Wiley & Sons, 2001.

[8] G. S. Fishman, *Monte Carlo: Concepts, algorithms, and applications*, Springer-Verlag, 1996.

[9] C. J. Geyer, "On the convergence of Monte Carlo maximum likelihood calculations," *Journal of the Royal Statistical Society. Series B (Methodological)*,

vol. 56, pp. 261–274, 1994.

[10] M. U. Gutmann and J. Hirayama, "Bregman divergence as general framework to estimate unnormalized statistical models," *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI)*, pp. 283–290, 2011.

[11] M. U. Gutmann and A. Hyvärinen, "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics," *Journal of Machine Learning Research*, vol. 13, pp. 307–361, 2012.

[12] M. U. Gutmann and A. Hyvärinen, "Estimation of unnormalized statistical models without numerical integration," *Proc. Workshop on Information Theoretic Methods in Science and Engineering (WITMSE2013)*, 2013.

[13] M. U. Gutmann and A. Hyvärinen, "A three-layer model of natural image statistics," *Journal of Physiology-Paris*, 2013, in press.

[14] S. Haykin, *Neural networks and learning machines*, 3rd ed., Prentice Hall, 2008.

[15] D. Koller, N. Friedman, L. Getoor, and B. Taskar, *Introduction to Statistical Relational Learning*, chapter Graphical Models in a Nutshell, pp. 13–55, MIT Press, 2007.

[16] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, pp. 1771–1800, 2002.

[17] A. Hyvärinen, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning Research*, vol. 6, pp. 695–709, 2005.

[18] A. Hyvärinen, "Some extensions of score matching," *Computational Statistics & Data Analysis*, vol. 51, pp. 2499–2512, 2007.

[19] E. L. Lehmann, and G. Casella, *Theory of Point Estimation*, 2nd ed., Springer, 1998.

[20] S. Z. Li, *Markov Random Field Modeling in Image Analysis*, Springer, 2009.

[21] M. Parry, A. P. Dawid, and S. Lauritzen, "Proper local scoring rules," *Annals of Statistics*, vol. 40, pp. 561–592, 2012.

[22] S. Lyu, "Interpretation and generalization of score matching," *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI)*, pp. 359–366, 2009.

[23] J. R. Movellan, "A minimum velocity approach to learning," *Machine Perception Laboratory Technical Report*, 2007.

[24] M. Pihlaja, M. U. Gutmann, and A. Hyvärinen, "A family of computationally efficient and simple estimators for unnormalized statistical models," *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI)*, pp. 442–449, 2010.

[25] N. L. Roux and Y. Bengio, "Representational power of restricted Boltzmann machines and deep belief networks," *Neural Computation*, vol. 20, pp. 1631–1649, 2008.

[26] J. Sohl-Dickstein, P. Battaglino, and M. DeWeese, "Minimum probability flow learning," *Proc. of Int. Conf. on Machine Learning*, pp. 905–912, 2011.

[27] M. Sugiyama, M. Kawanabe, and P. L. Chui, "Dimensionality reduction for density ratio estimation in high-dimensional spaces," *Neural Networks*, vol. 23, pp. 44–59, 2010.

[28] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density Ratio Estimation in Machine Learning*, Cambridge University Press, 2012.

[29] T. Tieleman, "Training restricted Boltzmann machines using approximations to the likelihood gradient," *Proc. of Int. Conf. on Machine Learning*, pp. 1064–1071, 2008.

[30] P. Vincent "A connection between score matching and denoising autoencoders," Neural Computation, vol. 23, pp. 1661–1674, 2011.