A Unified Statistically Efficient Estimation Framework for Unnormalized Models

Masatoshi Uehara
Harvard University

Takafumi Kanamori
Tokyo Institute of Technology
RIKEN AIP

Takashi Takneouchi
Future University Hakodate
RIKEN AIP

Takeru Matsuda
The University of Tokyo
RIKEN Center for Brain Science

Abstract

The parameter estimation of unnormalized models is a challenging problem. The maximum likelihood estimation (MLE) is computationally infeasible for these models since normalizing constants are not explicitly calculated. Although some consistent estimators have been proposed earlier, the problem of statistical efficiency remains. In this study, we propose a unified, statistically efficient estimation framework for unnormalized models and several efficient estimators, whose asymptotic variance is the same as the MLE. The computational cost of these estimators is also reasonable and they can be employed whether the sample space is discrete or continuous. The loss functions of the proposed estimators are derived by combining the following two methods: (1) density-ratio matching using Bregman divergence, and (2) plugging-in nonparametric estimators. We also analyze the properties of the proposed estimators when the unnormalized models are misspecified. The experimental results demonstrate the advantages of our method over existing approaches.

1 INTRODUCTION

Unnormalized models are widely used in many settings: Markov random fields (Besag, 1975), Boltzmann machines (Hinton, 2002), models in the independent component analysis (Hyvärinen, 2001), submodular diversity models (Tschiatschek et al., 2016) and generalized gamma distributions (Stacy, 1962). When the parametric model is denoted as $p(x;\theta)$, $p(x;\theta)$ is called an unnormalized (aka intractable) model if its normalizing constant $\int p(x;\theta)\mathrm{d}\mu(x)$ cannot be explicitly calculated, or it is difficult to compute in practice. For example, when μ is a counting measure as in the case of Markov random fields and Boltzmann machines, the computational cost exponentially increases with the dimension of the sample space. When μ is a Lebesgue measure, as in the case of the models in independent component analysis or generalized gamma distributions, this cannot be analytically calculated . When we use unnormalized models, we believe that the true data generating process belongs to the family $\{p(x;\theta)/\int p(x;\theta)\mathrm{d}\mu(x), \theta \in \Theta\}$, where Θ denotes a parameter space. Unnormalized models $p(x;\theta)$ can be converted to normalized models by dividing their normalizing constants; however, their explicit form cannot be obtained; therefore, an exact maximum likelihood estimation (MLE) is infeasible.

Several approaches for the estimation of unnormalized models have been suggested. Roughly, there are two major approaches. First, noise contrastive estimation (NCE) (Pihlaja et al., 2010; Gutmann and Hyvärinen, 2012; Hyvarinen and Morioka, 2016; Matsuda and Hyvärinen, 2019) and contrastive divergence (CD) (Hinton, 2002) rely on sampling techniques, such as importance sampling and Markov Chain Monte Carlo. Second, score matching (Hyvärinen, 2005; Hyvärinen, 2007; Dawid et al., 2012; Parry et al., 2012) and pseudo likelihood (Besag, 1975; Varin et al., 2011; Lindsay et al., 2011) use a tractable form without the aid of a sampling technique. The first approach is generally superior to the second approach in terms of statistical efficiency, whereas the second approach is superior to the first approach in terms of computational efficiency, leaving a tradeoff between computational and statistical efficiency.

In the present study, we propose a unified framework for the statistically efficient estimation of unnormalized models irrespective of whether the sample space is discrete or continuous. The estimators are defined as a form of M-estimators (van der Vaart, 1998) and their loss functions are derived by combining two methods: (1) density-ratio matching using Bregman divergence, and (2) plugging-in nonparametric estimators. These estimators are statistically efficient in the sense that the asymptotic variance is the same as that of the MLE; thus, the proposed estima-

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

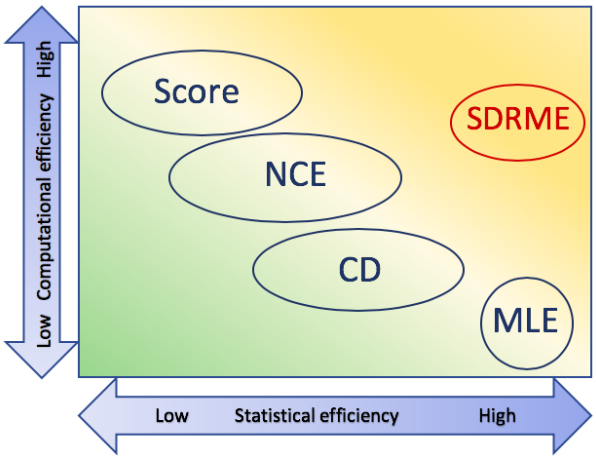

Figure 1: Comparison of methods, where Score stands for score matching; NCE stands for noise contrastive estimation, CD stands for contrastive divergence method; and a self density-ratio matching estimator (SDRME) is the proposed estimator. Note that statistically efficient estimators can be constructed in the case of NCE and CD. When the ratio of the auxiliary and original sample sizes is infinite in the NCE, the estimator becomes statistically efficient. However, implementing it is in practice infeasible. The same argument applies to the CD.

tors are superior to other previously proposed estimators in terms of statistical efficiency. Moreover, the proposed estimators do not rely on any sampling techniques and the evaluation cost of the objective function is $\mathcal{O}(n)$; therefore, they are competitive in terms of computational efficiency. Figure 1 illustrates a comparison of our proposed estimators to the other previously proposed estimators. To the best of our knowledge, the proposed estimators are the first statistically efficient estimators with evaluation cost $\mathcal{O}(n)$, which works in the continuous sample space.

Note that for a discrete sample space, Takenouchi and Kanamori (2017) proposed an efficient estimator, that can be seen as a special case from our proposed framework. Importantly, it is extended to the case of a continuous sample space based on the proposed framework.

2 PRELIMINARIES

Our general setting is as follows. Let us consider a situation in which an unnormalized model $p(x;\theta)$ is used, that is, for each $\theta \in \Theta$, $p(x;\theta)$ is a non-negative function and the normalizing constant defined by the integral $\int_{\mathcal{X}} p(x;\theta)\mathrm{d}\mu(x)$, is finite. The measure μ over the sample space $\mathcal{X}$ is a counting measure when the sample space is discrete, and a Lebesgue measure when the sample space is continuous. We refer to it herein as a baseline measure. We introduce a one-parameter extended model defined by $q(x;\tau) \equiv \exp(-c)p(x;\theta)$, $\tau \equiv (c, \theta^\top)^\top$ where c is also regarded as a parameter.

Our aim is to estimate θ using a set of identically independent distributed (i.i.d) n samples $\{x_i\}_{i=1}^n$ by assuming that these samples are obtained from the true distribution F_{η^*} with density $\eta^*(x)$ with respect to the baseline measure μ. Unless otherwise noted, we assume that the unnormalized model is well-specified, that is, there exists θ^* satisfying $\eta^* = \exp(-c^*)p(x;\theta^*)$, $\exp(c^*) = \int p(x;\theta^*)\mathrm{d}\mu(x)$. The problem of unnormalized models arises because it is extremely difficult or infeasible to calculate the normalizing constant analytically. In such a case, one should avoid a direct computation of the normalizing constant; therefore, the loss function of the MLE cannot be used. In this section, we review the Bregman divergence and the generalized NCE, needed to understand the proposed methods.

We summarize frequently used notations. We denote $\mathrm{E}_*(\cdot)$ as an expectation under the true density $\eta^*(x)$. The notations $\mathrm{Var}_*(\cdot)$ and $\tilde{\mathrm{E}}_*(\cdot)$ represent variance and empirical analogues. Notation $\mathbb{P}_n$ denotes an empirical distribution of n samples from the true distribution F_{η^*}. We denote $\mathrm{d}\mathbb{P}_n/\mathrm{d}\mu$ as p_n, evaluation at τ, i.e., $|_{\tau=\tau^*}$ as $|_{\tau^*}$, and ∇_x as the differentiation with respect to x. A summary of the notation is provided in a table in the Appendix A.

2.1 Bregman Divergence

Let $\mathbb{R}_{\geq 0}$ be a set of non-negative real numbers. We define $\mathcal{F}$ as a collection of non-negative real-valued functions on the sample space $\mathcal{X}$, and assume that $\mathcal{F}$ is a convex set. Given a convex function $\psi(u)$ on $\mathcal{F}$, the Bregman divergence (Bregman, 1967; Gneiting and Raftery, 2007; Dawid and Musio, 2014) on $\mathcal{F} \times \mathcal{F}$ is defined as $B_\psi(u, v) = \psi(u) - \psi(v) - \nabla\psi(v)(u - v)$, where $\nabla\psi(v)$ is a linear operator defined by $\lim_{\varepsilon\to+0}[\{\psi(v + \varepsilon h) - \psi(v)\}/\varepsilon] = \nabla\psi(v)(h)$. Here, h is a function on $\mathcal{X}$ such that $v+\varepsilon h \in \mathcal{F}$ holds for an arbitrary small $\varepsilon > 0$. The convexity of $\psi(u)$ guarantees the non-negativity of the Bregman divergence. We introduce two kinds of Bregman divergences; one is separable, while the other is non-separable.

The separable Bregman divergence is defined using the function $\psi(u)$:

$$\psi(u) = \mathrm{E}_*[f\{u(x)\}], \tag{1}$$

where $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a strictly convex function. For the differentiable f, the corresponding Bregman divergence $B_f(u, v; \eta^*)$ between u and v is given as

$$\mathrm{E}_*[f\{u(x)\} - f\{v(x)\} - f'\{v(x)\}\{u(x) - v(x)\}]. \tag{2}$$

For the strictly convex function f, the corresponding $B_f(u, v; \eta^*)$ vanishes if and only if $u = v$ up to a null set with respect to measure $\eta^*(x)\mathrm{d}\mu(x)$.

Example 2.1 *For $f(x) = 2x\log x - 2(1+x)\log(1+x)$, the corresponding $B_f(u, v; \eta^*)$ is known as the Jensen-Shannon divergence. In other cases, for $f(x) = x\log x$, we have the Kullback-Liber (KL) divergence. For $f(x) = x^m/\{m(m-1)\}$, we obtain the β-divergence (Basu et al., 1998; Murata et al., 2004).*

The Bregman divergence is non-separable if the convex function $\psi(u)$ is not expressed as (1). The pseudo-spherical divergence and the γ-divergence are examples of non-separable Bregman divergences, commonly used in robust inference (Kanamori and Fujisawa, 2014). The function $\psi(u)$ of the pseudo-spherical divergence is the γ-norm with $\gamma > 1$ under the density $\eta^*(x)$, that is, $\|u\|_\gamma = \mathrm{E}_*\{u(x)^{1+\gamma}\}^{-\gamma/(1+\gamma)}$. The pseudo-spherical divergence $B_{\mathrm{ps}}(u, v; \eta^*)$ is defined as follows:

$$\|u\|_\gamma - \frac{1}{\|v\|_\gamma^{\gamma-1}}\mathrm{E}_*\{v(x)^{\gamma-1}u(x)\}. \tag{3}$$

The pseudo-spherical divergence $B_{\mathrm{ps}}(u, v; \eta^*)$ vanishes if and only if u and v are linearly dependent. When we apply a log-transformation to each term in (3), this becomes a γ-divergence (Fujisawa and Eguchi, 2008), represented as

$$B_\gamma(u, v; \eta^*) = \frac{1}{\gamma}\log \mathrm{E}_*\{u(x)^\gamma\} + \frac{\gamma-1}{\gamma}\log \mathrm{E}_*\{v(x)^\gamma\} - \log \mathrm{E}_*\{v(x)^{\gamma-1}u(x)\}. \tag{4}$$

2.2 Generalized Noise Contrastive Estimation

We review an estimation method for unnormalized models focusing on a generalized NCE (Pihlaja et al., 2010; Gutmann and Hirayama, 2011). The strategy to estimate θ, c in $q(x;\tau)$ is matching a density ratio $q(x;\tau)/a(x)$ with a true densith ratio $q(x;\tau^*)/a(x)$, where $a(x)$ is a known auxiliary density, by generating samples from the distribution with a density $a(x)$.

Using a set of samples $\{y_i\}_{i=1}^n$ from the auxiliary distribution with a density $a(y)$ with respect to the baseline measure μ, the estimator $\hat{\tau}_{\mathrm{NC}}$ for τ is defined as the minimizer of the following function

$$\frac{1}{n}\sum_{i=1}^n r_{q,a}(y_i;\tau) f'\{r_{q,a}(y_i;\tau)\} - f\{r_{q,a}(y_i;\tau)\} - f'\{r_{q,a}(x_i;\tau)\}, \tag{5}$$

where $r_{q,a}(x;\tau) = q(x;\tau)/a(x)$, $f(x)$ is a strictly convex function, and the support of density $a(x)$ includes the support of $p(x;\theta)$. This estimation is derived from a divergence perspective as follows: let the divergence between the true distribution $\eta^*(x)$ and the one-parameter extended model $q(x;\tau)$ be $B_f\{r_{\eta^*,a}(x), r_{q,a}(x); a(x)\}$ when $r_{\eta^*,a}(x) = \eta^*(x)/a(x)$. We have $B_f\{r_{\eta^*,a}(x), r_{q,a}(x); a(x)\} \geq 0$ and $B_f\{r_{\eta^*,a}(x), r_{q,a}(x); a(x)\} = 0 \Leftrightarrow \eta^*(x) = q(x;\tau)$. Therefore, the estimation problem of τ is reduced to a minimization problem of $B_f\{r_{\eta^*,a}(x), r_{q,a}(x); a(x)\}$ with respect to τ. By subtracting the term not associated with $q(x;\tau)$ from $B_f\{r_{\eta^*,a}(x), r_{q,a}(x); a(x)\}$, we obtain the term:

$$-\int f'\{r_{q,a}(x)\}\eta^*(x)\mathrm{d}\mu(x) + \int [f'\{r_{q,a}(x)\} r_{q,a}(x) - f\{r_{q,a}(x)\}] a(x)\mathrm{d}\mu(x).$$

The loss function of $\hat{\tau}_{\mathrm{NC}}$, (5), is constructed using an empirical approximation of this term.

Unless otherwise noted, we hereafter assume the following properties for $f(x)$:

Assumption 1 *Function $f : \mathbb{R}_+ \to \mathbb{R}$ satisfies the following three properties: strictly convex, third-order differentiable and $f''(1) = 1$.*

Among $f(x)$ satisfying the abovementioned conditions, the estimator when $f(x) = 2x\log x - 2(1+x)\log(1+x)$ is proven to be optimal from the perspective of asymptotic variance, irrespective of the auxiliary distribution (Uehara et al., 2018). In this case, the loss function of the estimator becomes:

$$-\frac{1}{n}\sum_{i=1}^n \log\frac{r_{q,a}(x_i;\tau)}{1+r_{q,a}(x_i;\tau)} - \frac{1}{n}\sum_{i=1}^n \log\frac{1}{1+r_{q,a}(y_i;\tau)}. \tag{6}$$

This loss function is identical to the original NCE (Gutmann and Hyvärinen, 2012). Although it satisfies some aforementioned optimality, the asymptotic variance of the estimator derived from the above loss function is larger than that of the MLE. We can also use another type of $f(x)$. For example, when $f(x) = x\log x$, ths loss function is

$$-\frac{1}{n}\sum_{i=1}^n \log r_{q,a}(x_i;\tau) + \frac{1}{n}\sum_{i=1}^n r_{q,a}(y_i;\tau). \tag{7}$$

This is reduced to the same from as the one of Monte Carlo MLE (Geyer, 1994):

$$-\frac{1}{n}\sum_{i=1}^n \log p(x_i;\theta) + \log\left\{\frac{1}{n}\sum_{i=1}^n \frac{p(y_i;\theta)}{a(y_i)}\right\} \tag{8}$$

by profiling out c beforehand. The asymptotic variance of the Monte Carlo MLE is lager than that of the NCE (Riou-Durand and Chopin, 2018).

We have so far assumed that the sample size of the auxiliary distribution goes to infinity at the same rate as the size of the true distribution when considering an asymptotic regime. In this case, generalized NCE is not statistically efficient. In contrast, when the sample size of the auxiliary density grows faster than the sample size of the true

distribution, it is easily proved that the Monte Carlo MLE is statistically efficient. However, this asymptotic regime is suggesting that the evaluation cost of the objective function is larger than $\mathcal{O}(n)$, which is the order when MLE can be done exactly. This is problematic because it requires much computational time. Throughout this paper, our goal is to find an efficient estimator such that the evaluation cost of the objective function is $\mathcal{O}(n)$.

3 ESTIMATION WITH SELF DENSITY-RATIO MATCHING

We propose two types of statistically efficient estimators with a reasonable computational time. Our key idea is to match the ratio of the unnormalized model and nonparametrically estimated density using Bregman divergence. We introduce an estimator based on a separable Bregman divergence. Then, we introduce an estimator based on a non-separable Bregman divergence.

3.1 Separable Case

We introduce an estimator, called self density-ratio matching estimator (SDRME) for τ as a form of M-estimators:

$$\hat{\tau}_{\mathrm{s}} = \operatorname*{arg\,min}_{\tau \in \Theta_\tau} \mathrm{B}_f[h_1\{w(x;\tau)\}, h_2\{w(x;\tau)\}; p_n], \quad (9)$$

where Θ_τ is a parameter space for τ, $w(x) = q(x;\tau)/\hat{\eta}_n(x)$, $\hat{\eta}_n(x)$ is the nonparametric estimator using an entire set of samples, $q(x;\tau)$ is a one-parameter extended model in Section 2.2, $p_n = \mathrm{d}\mathbb{P}_n/\mathrm{d}\mu$, and $h_1(x)$ and $h_2(x)$ are functions satisfying the conditions mentioned in the next paragraph. We introduce h_1, h_2 to generalize the result as much as possible. More specifically, the loss function is written as

$$\frac{1}{n}\sum_{i=1}^{n} B_f\{h_1(w_i), h_2(w_i)\}, \quad (10)$$

where $w_i = q(x_i;\tau)/\hat{\eta}_n(x_i)$. Importantly, it requires only sample order $\mathcal{O}(n)$ calculation.

When the baseline measure is a counting measure, we use an empirical distribution $p_n(x)$ as $\hat{\eta}_n(x)$, whereas when the baseline measure is a Lebesgue measure, we use a kernel density estimator as $\hat{\eta}_n(x)$. Three conditions for $h_1(x), h_2(x)$ are assumed herein.

Assumption 2 *Functions $h_1 : \mathbb{R}_+ \to \mathbb{R}$ and $h_2 : \mathbb{R}_+ \to \mathbb{R}$ must be (2I) monotonically second-order differentiable increasing functions, (2II) $h_1(x) = h_2(x) \iff x = 1$, and (2III) $h_1'(1) \neq h_2'(1)$.*

Condition (2II) is required for the identification, and (2III) is needed to state the asymptotic normality of the estimators.

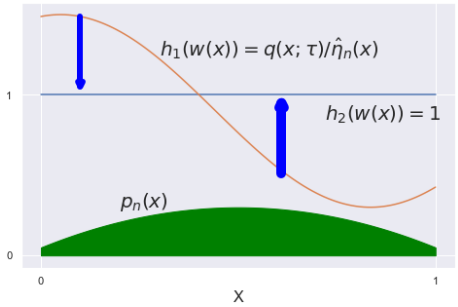

Figure 2: Graphical explanation of the SDRME with $h_1(w) = w$, $h_2(w) = 1$. The objective function is measuring the difference between $q(x;\tau)/\hat{\eta}_n(x)$ and $q(x;\tau^*)/\eta^*(x) = 1$ utilizing a Bregman divergence with density $p_n(x)$.

Table 1: Comparison between the generalized NCE and the SDRME in Example 3.1. Both methods are seen as matching a ratio with τ (ratio I) to a target ratio (ratio II). Here, efficiency means statistical efficiency.

	Generalized NCE	SDRME in Eg 3.1
Ratio I	$q(x;\tau)/a(x)$	$q(x;\tau)/\hat{\eta}_n(x)$
Ratio II	$q(x;\tau^*)/a(x)$	$q(x;\tau^*)/\eta^*(x) = 1$
Evaluation cost	$\mathcal{O}(n)$	$\mathcal{O}(n)$
Efficiency	No	Yes

This estimator works based on the following intuitive equivalence. By replacing $p_n(x)$ and $\hat{\eta}_n(x)$ with $\eta^*(x)$ in (9), we obtain $\mathrm{B}_f\{h_1(w), h_2(w); \eta^*\} = 0 \iff h_1(w) = h_2(w) \iff w = 1 \iff q(x;\tau) = \eta^*(x)$. As explained in Section 4, this estimator is rigorously proven to be consistent and efficient. Several specific choices can be considered as $h_1(w)$ and $h_2(w)$ as in Example 3.1. We explain the SDRME with a separable divergence in Figure 2 and Algorithm 1.

Example 3.1 (An extension of the generalized NCE) *Consider a case where $h_1(w) = w$ and $h_2(w) = 1$. We practically recommend this choice as $h_1(w)$ and $h_2(w)$. The loss function becomes*

$$\frac{1}{n}\sum_{i=1}^{n}\big\{ - f'(w_i) + w_i f'(w_i) - f(w_i)\big\}.$$

This is considered to be a natural extension of the generalized NCE as in Table 1 because the loss function (5) *is the same as the one above when we replace $a(x)$ with $\hat{\eta}_n(x)$, and y_i with x_i in* (5)*. Especially, when $f(x) = x\log x$, the loss functions corresponding* (7) *and* (8) *are*

$$-\frac{1}{n}\sum_{i=1}^{n}\log q(x_i;\tau) + \frac{1}{n}\sum_{i=1}^{n}\frac{q(x_i;\tau)}{\hat{\eta}_n(x_i)}, \quad (11)$$
$$-\frac{1}{n}\sum_{i=1}^{n}\log p(x_i;\theta) + \log\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{p(x_i;\theta)}{\hat{\eta}_n(x_i)}\right\}.$$

3.2 Non-separable Case

Similar to the separable Bregman divergence case, the pseudo-spherical divergence B_{ps} and the γ-divergence B_γ also provide statistically efficient estimators for unnormalized models. Following the analogy of the separable case when $h_1(w) = w^\alpha$, $h_2(w) = w^\beta$ $(\alpha \neq \beta)$, suppose that $B_{\mathrm{ps}}(w^\alpha, w^\beta; \eta^*) = 0$ holds. Then, w^α should be proportional to w^β because of the property of the pseudo-spherical divergence. In other words, $w(x)$ is a constant function. When $w(x;\theta)$ is $p(x;\theta)/\eta(x)$ and $\eta(x)$ is close to $\eta^*(x)$, $p(x;\theta)$ should be close to $\eta^*(x)$ up to the constant factor. This implies that the parameter θ can be estimated using the pseudo-spherical divergence. Replacing $\hat{\eta}_n(x)$ with a nonparametric estimator, the SDRME with the non-separable divergence $\hat{\theta}_{\text{ns-ps}}$ is obtained as

$$\arg\min_{\theta\in\Theta} B_{\mathrm{ps}}(w(x;\theta)^{\alpha/\gamma}, w(x;\theta)^{\beta/\gamma}; p_n),$$

and $w(x;\theta) = p(x;\theta)/\hat{\eta}_n(x)$, under the condition $\alpha \neq \beta$. Then, the loss function is

$$\left(\sum_{i=1}^n w_i^\alpha\right)^{\frac{1}{\gamma}} - \left(\sum_{i=1}^n w_i^\beta\right)^{(1-\gamma)/\gamma} \sum_{i=1}^n w_i^\delta,$$

where $\delta = (\alpha + \beta(\gamma-1))/\gamma, w_i = p(x_i;\theta)/\hat{\eta}_n(x_i)$. By taking a logarithm of each term as in (4), we can construct a loss function corresponding to γ-divergence. This is equal to $B_\gamma(w^\alpha, w^\beta; p_n)$:

$$\frac{1}{\gamma}\log\sum_{i=1}^n w_i^\alpha + \frac{\gamma-1}{\gamma}\log\sum_{i=1}^n w_i^\beta - \log\sum_{i=1}^n w_i^\delta. \quad (12)$$

We define estimator $\hat{\theta}_{\text{ns-}\gamma}$ as a minimizer of the above function with respect to θ over Θ. We explain the SDRME with a non-separable divergence in Algorithm 2.

Two things should be noted. First, compared with the case of the separable divergence, the unnormalized model $p(x;\theta)$ is directly used instead of a one-parameter extended model $q(x;\tau) = \exp(-c)p(x;\theta)$. This is due to the scale-invariance property of the pseudo-spherical divergence; $B_\gamma(u, v; p_n) = B_\gamma(u, \kappa v; p_n)$ for constant κ (Kanamori and Fujisawa, 2014, 2015). Second, when the baseline measure is a counting measure, Takenouchi and Kanamori (2017) proposed an estimator defined as a minimizer of the following function with respect to θ,

$$\frac{1}{\gamma}\log\sum_{x\in\mathcal{X}} c_x^{1-\alpha} p(x;\theta)^\alpha + \frac{\gamma-1}{\gamma}\log\sum_{x\in\mathcal{X}} c_x^{1-\beta} p(x;\theta)^\beta$$
$$-\log\sum_{x\in\mathcal{X}} c_x^{1-\delta} p(x;\theta)^\delta,$$

where $c_x = n_x/n$, and n_x is a sample number taking the value of x. This loss function is essentially the same as (12) by modifying the form of summing. The case was only considered when the sample space is discrete. However, it can be generalized to the case where the sample space is continuous, using our new unified perspective. For simplicity, hereafter, we assume $\delta = 0$ to eliminate the third term in (12). This restriction is also reasonable to obtain the convexity as seen in Appendix B.

Algorithm 1: SDRME with separable divergence

input : Data $\{x_i\}_{i=1}^n$ and model $p(x;\theta)$
output: $\hat{\tau}$
1 Set $h_1(x), h_2(x), f(x)$ (Default $h_1(x) = x$, $h_2(x) = 1, f(x) = x\log x$)
2 Make a nonparametric estimator $\hat{\eta}_n(x)$ from $\{x_i\}_{i=1}^n$
3 Define $w_i = q(x_i;\tau)/\hat{\eta}_n(x_i)$
4 Minimize (10) with respect to τ

Algorithm 2: SDRME with non-separable divergence

input : Data $\{x_i\}_{i=1}^n$ and model $p(x;\theta)$
output: $\hat{\theta}$
1 Set α, β, γ (Default $\alpha = -0.01, \beta = 0.99, \gamma = 1.01$)
2 Make a nonparametric estimator $\hat{\eta}_n(x)$ from $\{x_i\}_{i=1}^n$
3 Define $w_i = p(x_i;\theta)/\hat{\eta}_n(x_i)$
4 Minimize (12) with respect to θ

4 PROPERTIES OF SDRME

We prove that the asymptotic variance of estimators $\hat{\theta}_{\mathrm{s}}$ and $\hat{\theta}_{\text{ns-}\gamma}$ is identical to that of the MLE. We utilize the property in which our estimators take the form of the Z-estimators with infinite dimensional nuisance parameters (van der Vaart, 1998, 2002). For the proofs, refer to Appendix.

4.1 Efficiency in the Separable Case

First, we discuss the case when the divergence is separable. The estimator $\hat{\tau}_{\mathrm{s}}$ based on the separable divergence is defined as the minimizer of the following function $n^{-1}\sum_{i=1}^n B_f\{h_1(w_i), h_2(w_i)\}$, where $w_i = q(x_i;\tau)/\hat{\eta}_n(x_i)$ and $\hat{\eta}_n(x)$ is a nonparametric density estimator using an entire sample.

If $\hat{\eta}_n(x)$ was equal to $\eta^*(x)$, this estimator $\hat{\tau}_{\mathrm{s}}$ would be regarded as the solution to $\tilde{\mathrm{E}}_*[\phi(x;\tau,\eta^*)] = 0$, where $\phi(x;\tau,\eta)$ is

$$f[h_1\{w(x)\}] - f[h_2\{w(x)\}] -$$
$$f'[h_2\{w(x)\}]\,[h_1\{w(x)\} - h_2\{w(x)\}]\,,$$

and $w(x) = q(x;\tau)/\eta(x)$, by differentiating the loss function with respect to τ. Here, the moment condition $\mathrm{E}_*\{\phi(x;\tau,\eta^*)|_{\tau^*}\} = 0$ holds. This condition guarantees that the estimator converges to τ^*. However, this includes

the unknown term $\eta^*(x)$. By replacing $\eta^*(x)$ with the nonparametric estimator $\hat{\eta}_n(x)$, the estimator $\hat{\tau}_{\mathrm{s}}$ is still regarded as a Z-estimator. In fact, the estimator $\hat{\tau}_{\mathrm{s}}$ satisfies the equation $\tilde{\mathrm{E}}_*\{\phi(x;\tau,\hat{\eta}_n)\}=0$. The consistency holds as follows when the sample space is discrete. Before that, we assume the following conditions throughout this paper:

Assumption 3 *The model is $q(x;\tau)$ is C^2-function with respect to τ. The parameter space Θ_τ is compact and τ^* is in the interior of Θ_τ. The equation $q(x;\tau)=\eta(x)$ holds if and only if $\tau=\tau^*$.*

All of the conditions are common conditions used in MLE (van der Vaart, 1998).

Theorem 1 (Consistency in discrete space) $\hat{\tau}_s \xrightarrow{p} \tau^*$.

We next show the asymptotic normality of the estimator $\hat{\tau}_{\mathrm{s}}$ when the sample space is discrete.

Theorem 2 (Asymptotic normality in discrete space) *When the sample space is discrete, assume that (2a) the following matrix $\Omega = \mathrm{E}_*(\nabla_\tau \log q \nabla_{\tau^\top} \log q|_{\theta^*})$ is non-singular, and (2b) the second order derivative of the map $\eta \to \phi(x;\tau,\eta)$ is uniformly bounded around in a neighborhood of η^*. We then have:*

$$\sqrt{n}(\hat{\tau}_s-\tau^*)=\Omega^{-1}\mathbb{G}_n\left\{\nabla_\tau \log q(x;\tau)|_{\tau^*}\right\}+\mathrm{o}_p(1),$$
$$\sqrt{n}(\hat{\tau}_s-\tau^*)\xrightarrow{d}\mathcal{N}(0,\Omega^{-1}).$$

These assumptions originate from van der Vaart (2002, Theorem 6.18.). Assumption (2b) is required to control the remainder term in the proof. It is commonly used to state an asymptotic normality in MLE (van der Vaart, 1998).

The variance estimator for $\hat{\tau}_{\mathrm{s}}$ is easily constructed from Theorem 2. Finally, we prove that $\hat{\theta}_{\mathrm{s}}$ in $\hat{\tau}_{\mathrm{s}}=(\hat{c}_s,\hat{\theta}_{\mathrm{s}})$ is equivalent to MLE in terms of the asymptotic variance.

Corollary 1 *When the sample space is discrete, we have*

$$\sqrt{n}(\hat{\theta}_s-\theta^*)\xrightarrow{d}\mathcal{N}(0,\mathfrak{I}_{\theta^*}^{-1}),$$

where $\mathfrak{I}_{\theta^*}$ is the Fisher information matrix at θ^* of the normalized model, that is, $\mathrm{Var}_*\{S(x;\theta^*)\}$, where $S(x;\theta)=\nabla_\theta\left\{\log p(x;\theta)-\log\int p(x;\theta)\mathrm{d}\mu(x)\right\}$.

Next, we investigate the asymptotic behavior when the sample space is continuous. We use the kernel density estimator as a nonparametric estimator for $\eta^*(x)$. Note that any nonparametric estimators can also be applied. Assume that $\eta^*(x)$ belongs to a Hölder class of smoothness ν (Korostelev, 2011). The kernel density estimator is constructed as $\hat{\eta}_n(x)=(n\iota)^{-d_x}\sum_{i=1}^n K\left\{(x_i-x)/\iota\right\}$, where ι denotes a bandwidth, K denotes a d_x-dimensional kernel, and d_x denotes a dimension of x (Silverman, 1986). The overall error $\|\hat{\eta}_n-\eta^*\|_\infty$ is $\mathrm{O}_p((\log n/n)^{1/2}\iota^{-d_x/2}+\iota^\nu)$ by choosing high-order kernel (Fan and Hu, 1992). We have $\|\hat{\eta}_n-\eta^*\|_\infty=\mathrm{O}_p((\log n/n)^{-\frac{\nu}{2\nu+d_x}})$ by selecting the order of bandwidth correctly (Stones, 1982) .

From here, we analyze the asymptotic behavior of estimator $\hat{\tau}_{\mathrm{s}}$ when the sample space is continuous. We conclude that the estimator is still efficient.

Theorem 3 (Asymptotic normality in continuous space) *When the sample space is continuous, under the conditions used in Theorem 2 and (2c): $\nu/2>d_x$, (2d): $\int\|\nabla_\tau\log q(x;\tau)\|_{\tau^*}\mathrm{d}\mu(x)$ is finite, (2e): there is $\epsilon>0$ such that $\mathrm{E}_*\{\sup_{\|u\|<\epsilon}\|\nabla_\tau\log q(x+u;\tau)|_{\tau^*}\|^4\}<\infty$, then, $\hat{\tau}_s$ is consistent and*

$$\sqrt{n}(\hat{\tau}_s-\tau^*)\xrightarrow{d}\mathcal{N}(0,\Omega^{-1}),\ \sqrt{n}(\hat{\theta}_s-\theta^*)\xrightarrow{d}\mathcal{N}(0,\mathfrak{I}_{\theta^*}^{-1}),$$

where Ω is defined in Theorem 2.

Assumption (2c) is introduced to control a remainder term. In other words, this condition states that the convergence rate of $\hat{\eta}_n$ is $\mathrm{o}_p(n^{-1/4})$. This is a mild assumption to state an asymptotic normality such that the reminder term in the Taylor expansion is negligible. Assumptions (2d) and (2e) are introduced following Newey and Mcfadden (1994, Theorem 8.11)

4.2 Efficiency in the Non-separable Case

We consider an asymptotic analysis of estimator $\hat{\theta}_{\text{ns-}\gamma}$ with the γ-divergence. When μ is a counting measure, by differentiating (12) with respect to θ and multiplying by $-\gamma/\alpha$, we obtain $S_{\alpha,\beta}(x;\theta)$:

$$\int\frac{\{\nabla_\theta\log p(x;\theta)\}w(x;\theta)^\beta}{\int w(x;\theta)^\beta\mathrm{d}\mathbb{P}_{\mathrm{n}}(x)}\mathrm{d}\mathbb{P}_{\mathrm{n}}(x)-$$
$$\int\frac{\{\nabla_\theta\log p(x;\theta)\}w(x;\theta)^\alpha}{\int w(x;\theta)^\alpha\mathrm{d}\mathbb{P}_{\mathrm{n}}(x)}\mathrm{d}\mathbb{P}_{\mathrm{n}}(x),$$

where $w(x)=p(x;\theta)/\hat{\eta}_n(x)$. Importantly, compared with the case in Section 4.1, $p(x;\theta)$ is used in $w(x)$ instead of $q(x;\tau)$ because of the scale invariant property of γ-divergence. The estimator $\hat{\theta}_{\text{ns-}\gamma}$ satisfies $S_{\alpha,\beta}(x;\theta)=0$. The estimator $\hat{\theta}_{\text{ns-}\gamma}$ can be also seen as a Z-estimator with infinite and finite-dimensional nuisance parameters, that is, the solution to $\tilde{\mathrm{E}}_*[U_{\alpha,\beta}(x;\theta,c_1,c_2,\hat{\eta}_n)]=0$, where $U_{\alpha,\beta}(x;\theta,c_1,c_2,\eta)$:

$$\begin{bmatrix}\nabla_\theta\log p(x;\theta)\left\{\frac{p(x;\theta)^\beta}{\exp(c_1)}\eta(x)^{-\beta}-\frac{p(x;\theta)^\alpha}{\exp(c_2)}\eta(x)^{-\alpha}\right\}\\ \exp(c_1)-p(x;\theta)^\beta\eta(x)^{-\beta}\\ \exp(c_2)-p(x;\theta)^\alpha\eta(x)^{-\alpha}\end{bmatrix}.$$

The estimator validity is based on the moment condition $0=\mathrm{E}_*\{U_{\alpha,\beta}(x;\theta,c_1,c_2,\eta)|_{\theta^*,c_1^*,c_2^*,\eta^*}\}$, where $\exp(c_1^*)=\exp(c^*)^\beta$ and $\exp(c_2^*)=\exp(c^*)^\alpha$. Note that θ is a parameter of interests, and c_1, c_2, and η are nuisance parameters. We can derive the asymptotic results as in Section 4.1 and conclude that $\hat{\theta}_{\text{ns-}\gamma}$ is an efficient estimator.

Theorem 4 *When the sample space is discrete, under the conditions of Theorem 2, we have* $\sqrt{n}(\hat{\theta}_{ns\text{-}\gamma} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathfrak{J}_{\theta^*}^{-1})$. *When the sample space is continuous, under the conditions of Theorem 3, we have* $\sqrt{n}(\hat{\theta}_{ns\text{-}\gamma} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathfrak{J}_{\theta^*}^{-1})$.

Remark 4.1 *Refer to Appendix C.1 regarding the result when the models are mis-specified.*

4.3 Convexity

Convexity is important for optimization. We consider herein the convexity of loss functions. Suppose that the model is expressed by unnormalized exponential models, $q(x;\tau) = \exp(\tau^\top \xi(x))$, where $\xi(x)$ is a basis function and the corresponding basis function for c is -1. This model contains many types of unnormalized models such as Boltzmann machines and generalized gamma distributions. Regarding separable estimators $\hat{\tau}_{\mathrm{s}}$ in Example 3.1, we can find sufficient conditions to ensure the convexity of loss functions as Theorem 5. Refer to Appendix B for the specific examples of $f(x)$.

Theorem 5 *Suppose that* $f(z)$ *satisfies the inequality*

$$(2z-1)f''(z) + z(z-1)f'''(z) \geq 0$$

for arbitrary $z > 0$. *Then, the loss function of the estimator* $\hat{\tau}_s$ *in Example 3.1 is convex in* τ.

5 NUMERICAL EXPERIMENTS

We present herein several examples to illustrate the performance of the proposed procedure, and demonstrate that the asymptotic variance of the proposed estimators is the same as that of the MLE. We ran simulations in the settings of restricted Boltzmann machines, submodular diversity models, generalized gamma distributions. Regarding additional experiments using Poisson distributions, gamma distributions, refer to Appendix D.2.

We chose $h_1(x) = x$, $h_2(x) = 1$ as in Example 3.1. We used the following package for the kernel density estimation (Hayfield and Racine, 2008). We also used 6-th order kernel, and the bandwidth was selected by a cross validation based on the likelihood. We compare the following estimators:

- **MLE**: estimator by the MLE.
- **NCE**: estimator by the NCE (Gutmann and Hyvärinen, 2012). The sample size of the auxiliary distribution is set as the original sample size unless otherwise noted.
- **s-KL, s-Chi, s-JS**: proposed estimators, i.e., SDRME with a separable divergence $\hat{\theta}_{\mathrm{s}}$. When $f = x\log x$, denote **s-KL**. When $f = 0.5x^2$, denote **s-Chi**. When $f = 2x\log x - 2(1+x)\log(1+x)$, denote **s-JS**.
- **ns-γ**: SDRME with the non-separable γ-divergence, $\hat{\theta}_{\text{ns-}\gamma}$. Regarding the choice of α, β, see each section. We selected several α, β following an experiment section (Takenouchi and Kanamori, 2017)

Table 2: Monte Carlo mean of the KL divergence between the true and estimated densities scaled by sample size, $n\mathrm{KL}(\eta^*(\boldsymbol{v}), P(\boldsymbol{v}; \widehat{W}))$, in the RBM. Parenthesis indicates the standard deviation. The computational time (seconds) is measured per each iteration when $n = 4000$.

$\dim \boldsymbol{v} = 10, \dim \boldsymbol{h} = 2$, iteration: 20

n	s-KL	ns-γ	NCE	MLE
1000	12.8(3.56)	13.9(3.23)	19.7(5.31)	15.5(4.05)
2000	12.4(4.17)	13.1(5.13)	17.3(5.68)	12.8(4.35)
4000	14.3(5.77)	14.1(4.68)	18.0(7.35)	14.5(6.38)
Time	0.35	0.21	2.25	0.24

$\dim \boldsymbol{v} = 18, \dim \boldsymbol{h} = 2$, iteration: 20

n	s-KL	ns-γ	NCE	MLE
1000	18.7(4.03)	21.2(4.57)	76.7(87.9)	30.5(4.49)
2000	21.5(3.29)	23.0(3.34)	51.9(23.3)	31.2(5.70)
4000	25.9(8.97)	25.4(8.38)	38.8(9.25)	30.0(6.31)
Time	2.28	0.79	5.60	43.5

We do not compare the proposed estimators with score matching type estimators because the superiority of the statistical efficiency of NCE over score matching has been already shown (Gutmann and Hyvärinen, 2012).

5.1 Restricted Boltzmann Machine (RBM)

The RBM has parameter $W \in \mathbb{R}^{d_v \times d_h}$. The joint probability of the RBM with the visible nodes $\boldsymbol{v} \in \{+1,-1\}^{d_v}$ and hidden nodes $\boldsymbol{h} \in \{+1,-1\}^{d_h}$ is $P(\boldsymbol{v},\boldsymbol{h};W) \propto e^{\boldsymbol{v}^T W \boldsymbol{h}}$ and the marginal probability of $\boldsymbol{v}$ is $P(\boldsymbol{v};W) \propto \prod_{k=1}^{d_v} \cosh\{(\boldsymbol{v}^T W)_k\}$. The unnormalized model of the RBM is thus expressed as $q(\boldsymbol{v};\tau) = e^{-c}\prod_{k=1}^{d_v}\cosh\{(\boldsymbol{v}^T W)_k\}$ with parameter $\tau = (c, W)$.

We compared four estimators: **s-KL**, **ns-γ**, **NCE** and **MLE**. Regarding the results of **s-Chi**, **s-JS**, refer to Appendix D.2. The parameters in **ns-γ** were set to $\alpha = 0.01, \beta = -1$ and $\gamma = 1.01$. In low dimensional models, MLE is feasible because the normalized constant is accessible in practice. For **s-KL** and **ns-γ**, we incorporated the sample-based regularization to make the estimator stable. For the empirical distribution of the data $\hat{\eta}_n(\boldsymbol{v})$, the mixture model $(1-1/n)\hat{\eta}_n(\boldsymbol{v}) + u_n(\boldsymbol{v})/n$ was used as the non-parametric estimator of $\eta(\boldsymbol{v})$, where $u_n(\boldsymbol{v})$ is the empirical distribution of n samples generated from the uniform distribution over $\{+1,-1\}^{d_v}$. The additional term u_n is expected to work as a regularization. In the NCE, the

auxiliary distribution is defined as the uniform distribution, and the sample size from $a(y)$ is set to $5n$.

Table 2 shows Monte Carlo mean, and the standard deviation of the KL divergence between the true and estimated densities scaled by sample size. We confirm that the proposed methods, **s-KL** and **ns-γ**, are comparable to the MLE while they do not suffer from the computational burden of the normalization constant.[1] The accuracy of the NCE is lower than the other methods, and would be improved using larger samples from the auxiliary distribution, while the computational cost increases.

5.2 Submodular Diversity Model

Several types of probabilistic submodular models have been developed to model the diversity of item sets for applications such as recommendation systems and information summary. Among them, Tschiatschek et al. (2016) proposed the FLID (Facility LocatIon Diversity) model, which is a probability distribution over subsets S of $\{1, \cdots, V\}$. Specifically, FLID is defined as

$$P(S; u, w) \propto \exp\left\{\sum_{i \in S} u_i + \sum_{d=1}^{L} (\max_{i \in S} w_{i,d} - \sum_{i \in S} w_{i,d})\right\},$$

where u_i and $w_i = (w_{i,1}, \cdots, w_{i,L})$ represent the quality and latent embedding vector of the i-th item, respectively $(i = 1, \cdots, n)$. Since the computation of the normalization constant of FLID is prohibitive, Tschiatschek et al. (2016) proposed to estimate this model by using the NCE.

We compared **s-KL**, **ns-γ** and **NCE**. The parameters in **ns-γ** were set to $\alpha = -0.01, \beta = 0.99$ and $\gamma = 1.01$. We generated samples from the FLID model with $L = 2$ and $V = 12$. Each entry of u and w herein was sampled independently from the uniform distribution on $[0, 1]$. For the auxiliary distribution in the **NCE**, we used the product distribution following Tschiatschek et al. (2016).

Table 3 presents the Monte Carlo mean and standard error of the KL divergence between the true and estimated densities. The computation time of each estimator is also presnted. These results indicate the significant superiority of **s-KL** to **NCE** in terms of statistical efficiency with a reasonable computational time. We also observe that the performance of **s-KL** is more stable than that of **ns-γ**.

5.3 Generalized Gamma Distribution

We consider herein a distribution with the following unnormalized density $P(x; \theta_1, \theta_2) \propto \exp(-\theta_1 x^2) x^{\theta_2} \mathrm{I}(x > 0)$, when the baseline measure is the Lebesgue measure. This is referred to as a generalized gamma distribution (Stacy, 1962). We set the true value at $(\theta_1, \theta_2) = (1.3, 1.3)$.

We compared three estimators: **s-KL**, **ns-γ** and **NCE**. The parameters in **ns-γ** were set to $\alpha = -0.01, \beta = 0.99$ and $\gamma = 1.01$. Unlike Sections 5.1 and 5.2, we used a kernel density estimator for **s-KL** and **ns-γ**, and a half-normal distribution for **NCE** as an auxiliary distribution.

Table 4 presents the Monte Carlo mean of the mean squared errors. The result demonstrates the significant superiority of **s-KL** and **ns-γ** over **NCE** in terms of statistical efficiency with a reasonable computational time even when the sample space is continuous.

Table 3: Monte Carlo mean of the KL divergence between the true and estimated densities, scaled by sample size in a submodular diversity model. The computational time (seconds) is measured per iteration when $n = 2 \times 10^5$.

n	s-KL	ns-γ	NCE
5×10^4	36.4(7.3)	46.6(5.9)	44.4(4.0)
1×10^5	21.5(4.9)	46.4(4.2)	37.5(7.8)
2×10^5	16.9(7.6)	69.3(7.0)	35.9(20.9)
Time	4911	2020	9827

Table 4: Monte Carlo mean of the mean squared errors scaled by the sample size in a generalized gamma distribution. The computational time (seconds) is measured per iteration when $n = 2000$.

n	s-KL	ns-γ	NCE
500	68.2(10.3)	77.6(9.3)	250.3(64.0)
1000	67.9(7.7)	76.3(5.3)	240.7(69.7)
2000	68.3(5.4)	75.3(4.6)	246.1(43.5)
Time	1.3	1.3	0.5

6 CONCLUSION

We have proposed self density-ratio matching estimators. Importantly, proposed estimators are as statistically efficient as MLE without calculating normalizing constants, regardless of whether the sample space is discrete or continuous. In addition, they do not rely on any sampling techniques. Among the several estimators, we recommend using **s-KL** with $h_1(x) = x$, $h_2(x) = 1$ for practical purposes because its experimental performance is stable as shown in in Appendix D.2, its loss function is convex, and it is seen as a projection regarding the KL divergence, even when the model is misspecified. More extensive comparison is a future work.

[1] When $\boldsymbol{v} = 10$, MLE is fast because the calculation of the normalizing constants is easy. The computational problem in MLE arises when $\boldsymbol{v} = 18$.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments and suggestions.

Masatoshi Uehara was supported by MASASON Foundation. Takeru Matsuda was partially supported by JSPS KAKENHI Grant Numbers 16H06533 and 19K20220.

References

Basu, A., I. R. Harris, N. L. Hjort, and M. C. Jones (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika 85*, 549–559.

Besag, J. (1975). Statistical analysis of non-lattice data. *J. R. Statist. Soc. D 24*, 179–195.

Bregman, L. (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.

Dawid, A. and M. Musio (2014). Theory and applications of proper scoring rules. *METRON 72*, 169–183.

Dawid, A. P., S. Lauritzen, and M. Parry (2012). Proper local scoring rules on discrete sample spaces. *Ann. Statist. 40*, 593–608.

Fan, J. and T.-C. Hu (1992). Bias correction and higher order kernel functions. *Statist. Prob. Lett. 13*, 235–243.

Fujisawa, H. and S. Eguchi (2008). Robust parameter estimation with a small bias against heavy contamination. *J. Multivariate Anal. 99*, 2053–2081.

Geyer, C. (1994, January). On the convergence of monte carlo maximum likelihood calculations. *J. R. Statist. Soc. B 56*, 261–274.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Assoc. 102*, 359–378.

Gutmann, M. and J. Hirayama (2011). Bregman divergence as general framework to estimate unnormalized statistical models. *In Proccedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2011)*.

Gutmann, M. and A. Hyvärinen (2012). Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics. *J. Mach. Learn. Res. 13*, 307–361.

Hayfield, T. and J. S. Racine (2008). Nonparametric econometrics: The np package. *J. Statist.Software 27*.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput. 14*, 1771–1800.

Hyvärinen, A. (2001). *Independent component analysis*. New York: J. Wiley.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res. 6*, 695–709.

Hyvärinen, A. (2007). Some extensions of score matching. *Comput. Statist. Data Anal. 51*, 2499–2512.

Hyvarinen, A. and H. Morioka (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *In Advances in Neural Information Processing Systems (NIPS 2016)*.

Kanamori, T. and H. Fujisawa (2014). Affine invariant divergences associated with proper composite scoring rules and their applications. *Bernoulli 20*, 2278–2304.

Kanamori, T. and H. Fujisawa (2015). Robust estimation under heavy contamination using unnormalized models. *Biometrika 102*, 559–572.

Korostelev, Alexander; Korosteleva, O. (2011). *Mathematical Statistics : Asymptotic Minimax Theory*. Providence: American Mathematical Society.

Lindsay, B. G., G. Y. Yi, and J. Sun (2011). Issues and strategies in the selection of composite likelihoods. *Statist. Sinica 21*, 71–105.

Matsuda, T. and A. Hyvärinen (2019). Estimation of non-normalized mixture models. *In Proceedings of the 22nd International Workshop on Artificial Intelligence and Statistics (AISTATS)*.

Murata, N., T. Takenouchi, T. Kanamori, and S. Eguchi (2004). Information geometry of u-boost and bregman divergence. *Neural Comput. 16*, 1437–1481.

Newey, W. K. and D. Mcfadden (1994). Chapter 36 large sample estimation and hypothesis testing. In *Handbook of Econometrics*, Volume 4, pp. 2111–2245. Elsevier B.V.

Parry, M. F., A. P. Dawid, and S. L. Lauritzen (2012). Proper local scoring rules. *Ann. Statist. 40*, 561–592.

Pihlaja, M., M. Gutmann, and A. Hyvärinen (2010). A family of computationally efficient and simple estimators for unnormalized statistical models. *In Proccedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2010)*.

Riou-Durand, L. and N. Chopin (2018). Noise contrastive estimation: Asymptotic properties, formal comparison with mc-mle. *Electron. J. Statist. 12*, 3473–3518.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability. London ; New York: Chapman and Hall.

Stacy, E. (1962). A generalization of the gamma distribution. *Ann. Math. Statist. 33*, 1187–1192.

Stones, C. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist. 10*, 1040–1053.

Takenouchi, T. and T. Kanamori (2017). Statistical inference with unnormalized discrete models and localized homogeneous divergences. *J. Mach. Learn. Res. 18*, 1–26.

Tschiatschek, S., J. Djolonga, and A. Krause (2016). Learning probabilistic submodular diversity models via noise contrastive estimation. *In Proceedings of the 19th International Workshop on Artificial Intelligence and Statistics (AISTATS)*.

Uehara, M., T. Matsuda, and H. Komaki (2018). Analysis of noise contrastive estimation from the perspective of asymptotic variance. *arXiv preprint arXiv:1808.07983*.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge, UK ; New York, NY, USA: Cambridge University Press.

van der Vaart, A. W. (2002). *Semiparametric Statistics*. Lecture Notes in Mathematics ; 1781. Berlin, Heidelberg: Springer Berlin Heidelberg.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statist. Sinica 21*, 5–42.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1–25.