Parameter Estimation with Generalized Empirical Localization

Takashi Takenouchi(✉)

Future University Hakodate, RIKEN Center for Advanced Intelligence Project (AIP),
Hakodate, Japan
ttakashi@fun.ac.jp

Abstract. It is often difficult to estimate parameters of discrete models because of the computational cost for calculation of normalization constant, which enforces the model to be probability. In this paper, we consider a computationally feasible estimator for discrete probabilistic models using a concept of generalized empirical localization, which corresponds to the generalized mean of distributions and homogeneous γ-divergence. The proposed estimator does not require the calculation of the normalization constant and is asymptotically efficient.

Keywords: Unnormalized model · Asymptotic efficiency · γ-divergence

1 Introduction

In this paper, we focus on a problem of parameter estimation of discrete probabilistic models. A typical way for the estimation is the Maximum Likelihood Estimation (MLE) and the MLE is a "good" estimator which asymptotically satisfies the Cramér-Rao bound and is asymptotically efficient. In general, explicit solutions for the MLE cannot be obtained and then gradient-based optimization methods is usually required. But the calculation of the gradient includes the calculation of the normalization constant which makes the model to be in the probability space, and the calculation of the normalization constant is sometimes computationally intractable when the model is in a high-dimensional space. A typical example is the Boltzmann machine on $\mathcal{X} = \{+1, -1\}^p$,

$$\frac{\exp\left(\boldsymbol{\theta}_1 \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T \boldsymbol{\theta}_2 \boldsymbol{x}\right)}{\sum_{\boldsymbol{x}\in\mathcal{X}} \exp\left(\boldsymbol{\theta}_1 \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T \boldsymbol{\theta}_2 \boldsymbol{x}\right)} \tag{1}$$

and a calculation of the normalization constant of requires 2^p summation, which is hard to calculate as p is large. Other estimators derived from minimization of divergence measures [2] also suffer the computational problem. To tackle with the

This work was supported by JSPS KAKENHI Grant Numbers 16K00051 from MEXT, Japan.

F. Nielsen and F. Barbaresco (Eds.): GSI 2019, LNCS 11712, pp. 368–376, 2019.
https://doi.org/10.1007/978-3-030-26980-7_38

problem associated with the normalization constant, various kinds of approaches have been researched. Some methods are based on the Markov Chain Monte Carlo (MCMC) sampling and the contrastive divergence [7] is a well-known example. Another approach approximate the targeted probabilistic model by a tractable model by the mean-field approximation assuming independence of variables [11]. In this paper, we focus on an approach which considers an unnormalized model rather than the (normalized) probabilistic model. [8] defines information of "neighbor" by contrasting probability with that of a flipped variable and makes it possible to omit the calculation of normalization constant. [4] proposed a generalized local scoring rules on discrete sample spaces and [6] avoids the calculation of the normalization constant using a trick with auxiliary examples. [13] proposes an asymptotically efficient estimator without the calculation of the normalization constant, which consists of a concept of empirical localization and a homogeneous γ-divergence [5,10]. In this paper, we extend the concept of the empirical localization and propose a novel estimator which does not require the calculation of normalization constant. We investigate statistical properties of the proposed estimator and verify its validity with small experiments.

2 Settings

Let $\boldsymbol{x}$ be a d-dimensional vector in discrete space $\mathcal{X}$ such as $\{+1,-1\}^d$ or $\{1,2,\ldots\}^d$, and a bracket $\langle f\rangle$ for a function f on $\mathcal{X}$ denotes a sum of f over $\mathcal{X}$, $\langle f\rangle = \sum_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$. For a given dataset $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$, the empirical distribution $\tilde{p}(\boldsymbol{x})$ is defined as

$$\tilde{p}(\boldsymbol{x}) = \begin{cases} \frac{n_{\boldsymbol{x}}}{n} & \boldsymbol{x} \text{ is observed,} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $n_{\boldsymbol{x}}$ is number of examples $\boldsymbol{x}$ is observed. We consider a probabilistic model

$$\bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})}{Z_{\boldsymbol{\theta}}} \tag{3}$$

where $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ is a unnormalized model expressed as

$$q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp(\psi_{\boldsymbol{\theta}}(\boldsymbol{x})) \tag{4}$$

with a function $\psi_{\boldsymbol{\theta}}(\boldsymbol{x})$ parameterized by $\boldsymbol{\theta}$ and $Z_{\boldsymbol{\theta}}$ is a normalization constant defined as $Z_{\boldsymbol{\theta}} = \langle q_{\boldsymbol{\theta}}\rangle$ which enforces the (3) to be a probability function. Note that the unnormalized model (4) is not a probability function and $\langle q_{\boldsymbol{\theta}}\rangle = 1$ does not hold in general, and calculation of the normalization constant $Z_{\boldsymbol{\theta}}$ often requires a high computational cost. Then calculation of the Maximum Likelihood Estimator (MLE)

$$\hat{\boldsymbol{\theta}}_{MLE} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \sum_{i=1}^{n} \log \bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \tag{5}$$

or maximization process of the log-likelihood using its gradient

$$\sum_{i=1}^{n} \{\psi_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - \langle \bar{q}_{\boldsymbol{\theta}} \psi_{\boldsymbol{\theta}} \rangle\} \tag{6}$$

involves difficulty of computational cost derived from $Z_{\boldsymbol{\theta}}$. To overcome the difficulty of the computation of $Z_{\boldsymbol{\theta}}$, we consider combination of the γ-divergence and generalized mixture model.

2.1 γ-divergence

For two positive measure f, g, the γ-divergence [5] is defined as follows.

$$D_{\gamma}(f, g) = \frac{1}{1+\gamma} \log \left\langle f^{\gamma+1} \right\rangle + \frac{\gamma}{1+\gamma} \log \left\langle g^{\gamma+1} \right\rangle - \log \langle f g^{\gamma} \rangle \tag{7}$$

where γ is a positive constant. Note that $D_{\gamma}(f, g)$ is non-negative and is said to be homogeneous divergence because $D_{\gamma}(f, g) = 0$ holds if and only if $f \propto g$, rather than $f = g$. In the limit of $\gamma \to 0$, the γ-divergence reduces to the usual KL-divergence,

$$\left\langle f \log \frac{f}{g} - f + g \right\rangle. \tag{8}$$

Note that a combination of the γ-divergence and the unnormalized model does not solve the problem of computational cost because a term $D_{\gamma}(\tilde{p}, q_{\boldsymbol{\theta}})$ includes $\left\langle q_{\boldsymbol{\theta}}^{\gamma+1} \right\rangle$ whose computation also requires the same order with the normalization constant $Z_{\boldsymbol{\theta}}$.

2.2 Empirical Localization

Firstly, we briefly introduce a concept of empirical localization of the (unnormalized) model $\bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x})$(or $q_{\boldsymbol{\theta}}(\boldsymbol{x})$) with the empirical distribution $\tilde{p}(\boldsymbol{x})$ [13]. The empirical localization is interpreted as a generalized mean of $q_{\boldsymbol{\theta}}$ and $\tilde{p}$, and lies in e-flat subspace [1] as

$$\tilde{r}_{\alpha,\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\tilde{p}(\boldsymbol{x})^{\alpha} \bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha}}{\left\langle \tilde{p}^{\alpha} \bar{q}_{\boldsymbol{\theta}}^{1-\alpha} \right\rangle} = \frac{\tilde{p}(\boldsymbol{x})^{\alpha} q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha}}{\left\langle \tilde{p}^{\alpha} q_{\boldsymbol{\theta}}^{1-\alpha} \right\rangle}. \tag{9}$$

Note that the normalization constant $Z_{\boldsymbol{\theta}}$ in $\bar{q}_{\boldsymbol{\theta}}$ is canceled out and $\tilde{r}_{\alpha,\boldsymbol{\theta}}(\boldsymbol{x})$ does not depend on $Z_{\boldsymbol{\theta}}$, except for $\alpha = 0$. Also note that the denominator $\left\langle \tilde{p}^{\alpha} \bar{q}^{1-\alpha} \right\rangle$ (or $\tilde{r}_{\alpha,\boldsymbol{\theta}}(\boldsymbol{x})$ itself) can be easily calculated because the empirical distribution $\tilde{p}(\boldsymbol{x})$ has some values only on observed $\boldsymbol{x}$ in the dataset and is always 0 on the unobserved subset of $\mathcal{X}$. This implies the model $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ is empirically localized to the observed subset of domain $\mathcal{X}$ of dataset $\mathcal{D}$ and we can ignore the unobserved subset of $\mathcal{X}$, which leads to a drastic reduction of computational cost. We observe $\tilde{r}_{0,\boldsymbol{\theta}}(\boldsymbol{x}) = \bar{q}(\boldsymbol{x})$ and $\tilde{r}_{1,\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{p}(\boldsymbol{x})$, and (9) connects the empirical distribution $\tilde{p}$ and the normalized model $\bar{q}$ with the parameter α.

2.3 Estimator by Homogeneous Divergence and Empirical Localization

In [13], an estimator which does not require calculation of the normalization constant, was proposed by combining (7) and (9). The estimator is defined as

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} D_\gamma(\tilde{r}_{\alpha,\boldsymbol{\theta}}^{1/(1+\gamma)}, \tilde{r}_{\alpha',\boldsymbol{\theta}}^{1/(1+\gamma)}) \tag{10}$$

$$D_\gamma(\tilde{r}_{\alpha,\boldsymbol{\theta}}^{1/(1+\gamma)}, \tilde{r}_{\alpha',\boldsymbol{\theta}}^{1/(1+\gamma)}) = \frac{1}{1+\gamma}\log\left\langle \tilde{p}^{\alpha} q_{\boldsymbol{\theta}}^{1-\alpha}\right\rangle + \frac{\gamma}{1+\gamma}\log\left\langle \tilde{p}^{\alpha'} q_{\boldsymbol{\theta}}^{1-\alpha'}\right\rangle - \log\left\langle \tilde{p}^{\beta} q_{\boldsymbol{\theta}}^{1-\beta}\right\rangle \tag{11}$$

where $\alpha \neq \alpha'$ and $\beta = (\alpha + \gamma\alpha')/(1+\gamma)$. We observe that a setting with $\alpha = 1, \alpha' = 0$ and $\gamma \to 0$ corresponds to the conventional MLE. Note that the empirical risk (11) does not include the calculation of the normalization constant $Z_{\boldsymbol{\theta}}$ and can be easily calculated.

The estimator (10) has the following good statistical properties.

Proposition 1 ([13]). *Let us assume that $\psi_{\boldsymbol{\theta}}(\boldsymbol{x})$ is written as $\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x})$ with a fixed vector function $\boldsymbol{\psi}(\boldsymbol{x})$. Then the risk function (10) is convex with respect to $\boldsymbol{\theta}$ when $\beta = 1$ holds.*

Proposition 2 ([13]). *The estimator (10) is Fisher consistent and asymptotically efficient.*

3 Proposed Estimator

The empirical localization (9) can be interpreted as a generalized mean of a constant 1 and a distribution ratio $q_{\boldsymbol{\theta}}/\tilde{p}$, and is rewritten as

$$\begin{aligned}\tilde{r}_{\alpha,\boldsymbol{\theta}}(\boldsymbol{x}) \propto \tilde{p}(\boldsymbol{x})^{\alpha} q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha} &= \tilde{p}(\boldsymbol{x})\left(\frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})}{\tilde{p}(\boldsymbol{x})}\right)^{1-\alpha} \\ &= \tilde{p}(\boldsymbol{x})\exp\left(\alpha\log 1 + (1-\alpha)\log\frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})}{\tilde{p}(\boldsymbol{x})}\right).\end{aligned} \tag{12}$$

We can extend the concept of (12) to the quasi-arithmetic mean, with a monotonically increasing function u and its inverse function ξ, as follows.

$$\tilde{r}_{u,\alpha,\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{p}(\boldsymbol{x}) u\left(\alpha\xi(1) + (1-\alpha)\xi\left(\frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})}{\tilde{p}(\boldsymbol{x})}\right)\right). \tag{13}$$

By transforming the function $u(z)$ to $u(z-a)$, we can set $\xi(1) = 0$ without loss of generality. The generalized version of empirical localization (13) is rewritten as

$$\tilde{r}_{u,\alpha,\boldsymbol{\theta}}(\boldsymbol{x}) \propto \begin{cases} n_{\boldsymbol{x}} u\left((1-\alpha)\xi\left(\frac{n q_{\theta}(\boldsymbol{x})}{n_{x}}\right)\right) & \boldsymbol{x} \text{ is observed} \\ 0 & \text{otherwise,}\end{cases} \tag{14}$$

and the model can be easily calculated because we can omit the unobserved domain. We show two examples associated with β-divergence and η-divergence [3] which are employed for the purpose of robust estimation [9,12].

Example 1. For $u(z) = (1+\beta z)^{1/\beta}$ and $\xi(z) = \frac{z^\beta - 1}{\beta}$, we have

$$\tilde{r}_{u,\alpha,\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{p}(\boldsymbol{x}) \left((1-\alpha) \left(\frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})}{\tilde{p}(\boldsymbol{x})} \right)^{\beta} + \alpha \right)^{1/\beta} \tag{15}$$

Example 2. For $u(z) = (1+\eta)e^z - \eta$ and $\xi(z) = \log \frac{z+\eta}{1+\eta}$, we have

$$\tilde{r}_{u,\alpha,\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{p}(\boldsymbol{x}) \left\{ (1+\eta) \left(\frac{\frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})}{\tilde{p}(\boldsymbol{x})} + \eta}{1+\eta} \right)^{1-\alpha} - \eta \right\} \tag{16}$$

Example 3. For $u(z) = -\frac{1}{z}$ and $\xi(z) = -\frac{1}{z}$, we have

$$\tilde{r}_{u,\alpha,\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\tilde{p}(\boldsymbol{x}) q_{\boldsymbol{\theta}}(\boldsymbol{x})}{\alpha q_{\boldsymbol{\theta}}(\boldsymbol{x}) + (1-\alpha)\tilde{p}(\boldsymbol{x})} \tag{17}$$

We propose a novel estimator for discrete probabilistic model, which can be constructed without calculation of the normalization constant $Z_{\boldsymbol{\theta}}$. The proposed estimator is defined by combining the (13) and γ-divergence with two hyper-parameters $\alpha, \alpha' (\alpha \neq \alpha')$, as follows.

$$\hat{\boldsymbol{\theta}} = \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}} D_\gamma((\tilde{r}_{u,\alpha,\boldsymbol{\theta}})^{1/(1+\gamma)}, (\tilde{r}_{u,\alpha',\boldsymbol{\theta}})^{1/(1+\gamma)}) \tag{18}$$

Note that when $q_{\boldsymbol{\theta}}(\boldsymbol{x}) \propto \tilde{p}(\boldsymbol{x})$ holds, we observe that $\tilde{r}_{u,\alpha,\boldsymbol{\theta}}(\boldsymbol{x}) \propto \tilde{r}_{u,\alpha',\boldsymbol{\theta}}(\boldsymbol{x})$ and $D_\gamma((\tilde{r}_{u,\alpha,\boldsymbol{\theta}})^{1/(1+\gamma)}, (\tilde{r}_{u,\alpha',\boldsymbol{\theta}})^{1/(1+\gamma)}) = 0$ holds.

4 Statistical Property

In this section, we investigate statistical property of the proposed estimator. Firstly, we show the Fisher consistency of the proposed estimator.

Proposition 3. *Let $\boldsymbol{\theta}_0$ be a true parameter of the underlying distribution, i.e., $p(\boldsymbol{x}) = \bar{q}_{\boldsymbol{\theta}_0}(\boldsymbol{x})$. Then*

$$\boldsymbol{\theta}_0 = \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}} D_\gamma(r_{u,\alpha,\boldsymbol{\theta}}, r_{u,\alpha',\boldsymbol{\theta}}) \tag{19}$$

holds for arbitrary γ, α, $\alpha' (\alpha \neq \alpha')$ and $\boldsymbol{\theta}_0$.

Proof. The proposed estimator satisfies the equilibrium equation

$$0 = \left. \frac{\partial D_\gamma((\tilde{r}_{u,\alpha,\boldsymbol{\theta}})^{1/(1+\gamma)}, (\tilde{r}_{u,\alpha',\boldsymbol{\theta}})^{1/(1+\gamma)})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{20}$$

implying the Fisher consistency.

Secondly, we investigate the asymptotic distribution of the proposed estimator.

Proposition 4. *Let $\boldsymbol{\theta}_0$ be a true parameter of the underlying distribution. Then, under mild regularity condition, the proposed estimator asymptotically follows*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \mathcal{N}(\mathbf{0}, I(\boldsymbol{\theta}_0)^{-1}) \tag{21}$$

where $\mathcal{N}$ is the Normal distribution and $I(\boldsymbol{\theta}_0) = V_{\bar{q}_{\boldsymbol{\theta}_0}}[\psi'_{\boldsymbol{\theta}_0}]$ is the Fisher information matrix.

Proof. Let us assume that the empirical distribution is written as $\tilde{p}(\boldsymbol{x}) = \bar{q}_{\boldsymbol{\theta}_0}(\boldsymbol{x}) + \epsilon(\boldsymbol{x})$. By expanding the equilibrium condition (20) around $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\epsilon(\boldsymbol{x}) = 0$, we have

$$\begin{aligned} 0 \simeq & \left.\frac{\partial D_\gamma((\tilde{r}_{u,\alpha,\boldsymbol{\theta}})^{1/(1+\gamma)}, (\tilde{r}_{u,\alpha',\boldsymbol{\theta}})^{1/(1+\gamma)})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ & + \left.\frac{\partial^2 D_\gamma((\tilde{r}_{u,\alpha,\boldsymbol{\theta}})^{1/(1+\gamma)}, (\tilde{r}_{u,\alpha',\boldsymbol{\theta}})^{1/(1+\gamma)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0). \end{aligned} \tag{22}$$

Using the delta method [14], we have

$$\left.\frac{\partial D_\gamma((\tilde{r}_{u,\alpha,\boldsymbol{\theta}})^{1/(1+\gamma)}, (\tilde{r}_{u,\alpha',\boldsymbol{\theta}})^{1/(1+\gamma)})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} - \left.\frac{\partial D_\gamma((r_{u,\alpha,\boldsymbol{\theta}})^{1/(1+\gamma)}, (r_{u,\alpha',\boldsymbol{\theta}})^{1/(1+\gamma)})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \tag{23}$$

$$\simeq C \left\langle \psi'_{\boldsymbol{\theta}_0} \epsilon \right\rangle \tag{24}$$

where C is a constant, and from the central limit theorem, we observe $\sqrt{n}\left\langle \psi'_{\boldsymbol{\theta}_0} \epsilon \right\rangle$ asymptotically follows the normal distribution with mean 0 and variance $I(\boldsymbol{\theta}_0) = V_{\bar{q}_{\boldsymbol{\theta}_0}}[\psi'_{\boldsymbol{\theta}_0}]$. From the law of large number, we observe that the second term in the rhs of (22) converges to $-CI(\boldsymbol{\theta}_0)$ in the limit of $n \to \infty$, which concludes the proposition.

The asymptotic variance in (21) implies that the proposed estimator is asymptotically efficient and has the same efficiency with the MLE, which asymptotically attains the Cramér-Rao bound. Also note that the asymptotic variance of the proposed estimator does not depend on choice of α, α', γ.

5 Experiments

We numerically investigated properties of the proposed estimator with a small synthetic dataset. Let $\bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x})$ be a 5-dimensional Boltzmann machine

$$\bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\exp\left(\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\theta} \boldsymbol{x}\right)}{Z_{\boldsymbol{\theta}}} \tag{25}$$

whose parameter $\boldsymbol{\theta}$ follows the normal distribution with mean 0 and variance 1. We generated 20 sets of datasets including 4000 examples and compared the following method.

1. MLE: Maximum likelihood estimator
2. gamma: The proposed estimator with $u(z) = \exp(z)$ [13]
3. IS: The proposed estimator with $u(z) = -\frac{1}{z}$
4. eta: The proposed estimator with $u(z) = (1+\eta)\exp(z) - \eta$
5. beta: The proposed estimator with $u(z) = (1+\beta z)^{1/\beta}$

Figure 1(a) shows a box plot of MSEs of parameters, $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2$ in a logarithmic scale, with various deformation function u in (13). Figure 1(b) shows a box plot of computational times for each estimator in a logarithmic scale. We observe that some of the proposed estimator is comparable with the MLE, while the computational time of the proposed estimator is drastically reduced compared with that of the MLE.

A reason of why the proposed estimator with some functions u are inferior to the MLE is a shortage of examples. The theoretical result shown in Sect. 4 is based on assumptions of asymptotics and requires a lot of examples to assure

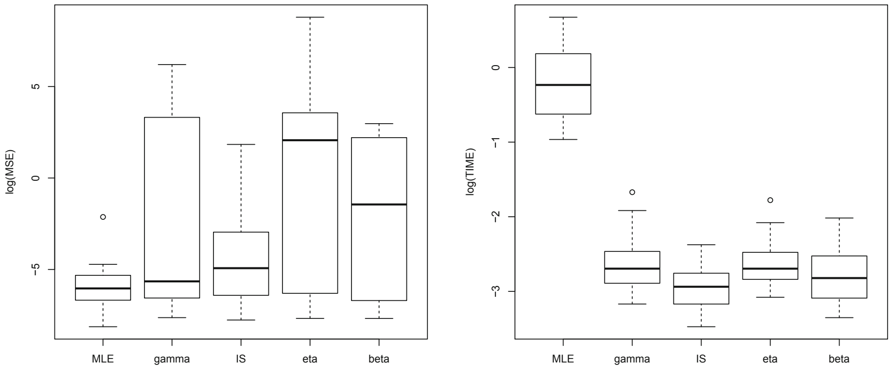

Fig. 1. $n = 4000$. (a) Box plot of estimation errors, $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2$ of each method. (b) Box plot of computational time of each method.

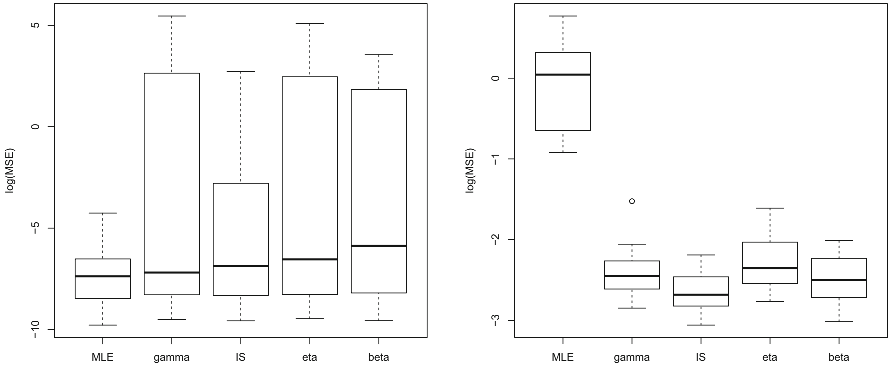

Fig. 2. $n = 16000$. (a) Box plot of estimation errors, $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2$ of each method. (b) Box plot of computational time of each method.

the asymptotic efficiency. We executed an another experiment with the same setting except for the number n of examples. Figure 2(a), (b) show results for $n = 16000$ and we observe that performance of the proposed estimator (IS, eta, beta) is improved at the same level as the MLE while required computational cost is still drastically fewer.

6 Conclusion

We proposed the novel estimator for discrete probabilistic model, which does not require calculation of the normalization constant. The proposed estimator is constructed by a combination of the γ-divergence and generalized empirical localization, which can be interpreted as the generalized mean of distributions. We investigated statistical properties of the proposed estimator and showed that the proposed estimator asymptotically has the same efficiency with the MLE and demonstrated the asymptotic efficiency with the small experiment.

References

1. Amari, S., Nagaoka, H.: Methods of Information Geometry, Translations of Mathematical Monographs, vol. 191. Oxford University Press, Oxford (2000)
2. Basu, A., Shioya, H., Park, C.: Statistical Inference: The Minimum Distance Approach. Chapman and Hall/CRC, Boca Raton (2011)
3. Cichocki, A., Cruces, S., Amari, S.i.: Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. Entropy **13**(1), 134–170 (2011)
4. Dawid, A.P., Lauritzen, S., Parry, M.: Proper local scoring rules on discrete sample spaces. Ann. Stat. **40**(1), 593–608 (2012)
5. Fujisawa, H., Eguchi, S.: Robust parameter estimation with a small bias against heavy contamination. J. Multivar. Anal. **99**(9), 2053–2081 (2008)
6. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 297–304 (2010)
7. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Comput. **14**(8), 1771–1800 (2002)
8. Hyvärinen, A.: Some extensions of score matching. Comput. Stat. Data Anal. **51**(5), 2499–2512 (2007)
9. Mihoko, M., Eguchi, S.: Robust blind source separation by beta divergence. Neural Comput. **14**(8), 1859–1886 (2002)
10. Nielsen, F., Nock, R.: Patch matching with polynomial exponential families and projective divergences. In: Amsaleg, L., Houle, M.E., Schubert, E. (eds.) SISAP 2016. LNCS, vol. 9939, pp. 109–116. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46759-7_8
11. Opper, M., Saad, D. (eds.): Advanced Mean Field Methods: Theory and Practice. MIT Press, Cambridge (2001)
12. Takenouchi, T., Eguchi, S.: Robustifying AdaBoost by adding the naive error rate. Neural Comput. **16**(4), 767–787 (2004)

13. Takenouchi, T., Kanamori, T.: Statistical inference with unnormalized discrete models and localized homogeneous divergences. J. Mach. Learn. Res. **18**(56), 1–26 (2017). http://jmlr.org/papers/v18/15-596.html
14. Van der Vaart, A.W.: Asymptotic Statistics. Cambridge University Press, Cambridge (1998)