Journal of Machine Learning Research 22 (2021) 1-33 Submitted 12/20; Revised 6/21; Published 7/21

Information criteria for non-normalized models

Takeru Matsuda TAKERU.MATSUDA@RIKEN.JP
RIKEN Center for Brain Science

Masatoshi Uehara MU223@CORNELL.EDU
Department of Computer Science, Cornell University

Aapo Hyvärinen AAPO.HYVARINEN@HELSINKI.FI
Department of Computer Science, University of Helsinki

Editor: Qiang Liu

Abstract

Many statistical models are given in the form of non-normalized densities with an intractable normalization constant. Since maximum likelihood estimation is computationally intensive for these models, several estimation methods have been developed which do not require explicit computation of the normalization constant, such as noise contrastive estimation (NCE) and score matching. However, model selection methods for general non-normalized models have not been proposed so far. In this study, we develop information criteria for non-normalized models estimated by NCE or score matching. They are approximately unbiased estimators of discrepancy measures for non-normalized models. Simulation results and applications to real data demonstrate that the proposed criteria enable selection of the appropriate non-normalized model in a data-driven manner.

Keywords: energy-based model, model selection, noise contrastive estimation, score matching

1. Introduction

We consider here the estimation of parametric statistical models which are non-normalized (energy-based). A non-normalized model is one in which the probability density does not integrate to unity. Given a functional form $\widetilde{p}(x \mid \theta)$ for the parametrized density[1], the actual normalized model would be given by

$$p(x \mid \theta) = \frac{1}{Z(\theta)} \widetilde{p}(x \mid \theta), \tag{1}$$

where

$$Z(\theta) = \int \widetilde{p}(x \mid \theta) \mathrm{d}x.$$

[1] In this paper, all measures are assumed to have densities (on $\mathbb{R}^d$ or its subset) and a measure and its density are identified without loss of generality for convenience. We explicitly write input arguments (such as x, y, z) of densities for clarity.

In the framework considered here, estimation of the parameters is attempted without computing the integral defining the normalization constant $Z(\theta)$, which is assumed to be too difficult to compute. Many statistical models naturally have such property: for instance, Markov random field models (Li, 2001), directional distributions (Mardia and Jupp, 2008; Chikuse, 2003), truncated Gaussian graphical models (Lin et al., 2016), network models (Caimo and Friel, 2011), and energy-based overcomplete independent component analysis models (Teh et al., 2004). Since maximum likelihood estimation is computationally intensive for such non-normalized models, several estimation methods have been developed which avoid calculation of the normalization constant. These methods include pseudo-likelihood (Besag, 1974), Monte Carlo maximum likelihood (Geyer, 1994), contrastive divergence (Hinton, 2002), score matching (Hyvärinen, 2005) and noise contrastive estimation (Gutmann and Hyvärinen, 2010). Among them, noise contrastive estimation (NCE) is applicable to general non-normalized models for both continuous and discrete data. In NCE, the normalization constant $Z(\theta)$ is estimated together with the unknown parameter θ by discriminating between data and artificially generated noise. On the other hand, score matching is a computationally efficient method for continuous data which is based on a trick of integration by parts. The idea of score matching has been generalized to the theory of proper local scoring rules (Parry et al., 2012) and also applied to Bayesian model selection with improper priors (Dawid and Musio, 2015; Shao et al., 2019). Several studies extended score matching to discrete data (Hyvärinen, 2007; Lyu, 2009). Recently, Stein's method has been applied to estimation of non-normalized models (Barp et al., 2019; Liu et al., 2019).

Although non-normalized models enable more flexible modeling of data-generating processes, information criteria-based model selection methods have not been proposed for NCE and score matching, to the best of our knowledge. In general, model selection is the task of selecting a statistical model from several candidates based on data (Burnham and Anderson, 2002; Claeskens and Hjort, 2008; Konishi and Kitagawa, 2008), where different candidates can have different number of parameters. By selecting an appropriate model in a data-driven manner, we obtain better understanding of the underlying phenomena and also better prediction of future observations. Akaike (1974) established a unified approach to model selection from the viewpoint of information theory and entropy. Specifically, he proposed Akaike Information Criterion (AIC) as a measure of the discrepancy between the true and estimated model in terms of the Kullback–Leibler divergence. Thus, the model with the minimum AIC is selected as the best model. AIC is widely used in many areas and has been extended by several studies (Takeuchi, 1976; Konishi and Kitagawa, 1996; Kitagawa, 1997; Spiegelhalter et al., 2002; Watanabe and Opper, 2010). However, these existing information criteria assume that the model is normalized and thus they are not applicable to non-normalized models.

In this study, we develop information criteria for non-normalized models estimated by NCE or score matching. For NCE, based on the observation that NCE is a projection with respect to a Bregman divergence (Gutmann and Hirayama, 2011), we propose noise contrastive information criterion (NCIC) as an approximately unbiased estimator of the model discrepancy induced by this Bregman divergence. Note that AIC (Akaike, 1974) was developed as an approximately unbiased estimator of the Kullback-Leibler discrepancy. Similarly, for score matching, we propose score matching information criterion (SMIC) as an approximately unbiased estimator of the model discrepancy induced by the Fisher divergence (Lyu, 2009). Thus, the non-normalized model with the minimum NCIC or SMIC is selected as the best model. Experimental results show that these procedures successfully select the appropriate non-normalized model in a data-driven manner. Therefore, this study increases the practicality of non-normalized models. Note that Ji and Seymour (1996) and Varin and Vidoni (2005) proposed information criteria based on the pseudo-likelihood and composite likelihood, respectively. Whereas their criteria are useful for discrete-valued data, our criteria are applicable to continuous-valued data, and NCIC is equally applicable to discrete-valued data.

This paper is organized as follows. In Sections 2 and 3, we briefly review noise contrastive estimation (NCE) and score matching, respectively. In Section 4, we review the theory of Akaike information criterion (AIC) and Takeuchi information criterion (TIC). In Sections 5 and 6, we develop information criteria for non-normalized models estimated by NCE and score matching, respectively. In Section 7, we confirm the validity of NCIC and SMIC by numerical experiments. In Section 8, we apply NCIC and SMIC to real data of natural image, RNAseq and wind direction. In Section 9, we discuss extension of NCIC to non-normalized mixture models. In Section 10, we give concluding remarks.

2. Noise contrastive estimation (NCE)

In this section, we briefly review noise contrastive estimation (NCE), which is a general method for estimating non-normalized models. For more detail, see Gutmann and Hyvärinen (2012).

2.1 Procedure of NCE

Suppose we have N i.i.d. samples $x^{(1)}, \dots, x^{(N)}$ from a parametric distribution (1). In NCE, we rewrite the non-normalized model (1) to

$$\log p(x \mid \theta, c) = \log \widetilde{p}(x \mid \theta) + c, \tag{2}$$

where $c = -\log Z(\theta)$. We regard c as an additional parameter and estimate it together with θ. Note that the final estimate $p(x \mid \hat{\theta}, \hat{c})$ is not normalized in general.

In addition to data $x^{(1)}, \dots, x^{(N)}$ from the non-normalized model (1), we generate M i.i.d. noise samples $y^{(1)}, \dots, y^{(M)}$ from a noise distribution with density $n(y)$. In practice, the noise distribution is usually chosen to be as close as possible to the true data distribution. For example, when the data is a random vector, the normal distribution with the same mean and covariance with data is often used as the noise distribution. Note that the noise distribution can be non-normalized itself, in which case MCMC can be employed for sampling $y^{(1)}, \dots, y^{(M)}$ (Riou-Durand and Chopin, 2018). Then, we estimate (θ, c) by discriminating between the data and noise as accurately as possible:

$$(\hat{\theta}_{\mathrm{NCE}}, \hat{c}_{\mathrm{NCE}}) = \arg\min_{\theta, c} \hat{d}_{\mathrm{NCE}}(\theta, c), \tag{3}$$

where

$$\hat{d}_{\mathrm{NCE}}(\theta, c) = -\frac{1}{N} \sum_{t=1}^{N} \log \frac{N p(x^{(t)} \mid \theta, c)}{N p(x^{(t)} \mid \theta, c) + M n(x^{(t)})} - \frac{1}{N} \sum_{t=1}^{M} \log \frac{M n(y^{(t)})}{N p(y^{(t)} \mid \theta, c) + M n(y^{(t)})}. \tag{4}$$

The objective function $\hat{d}_{\rm NCE}$ is the negative log-likelihood of the logistic regression classifier. In other words, each term in $\hat{d}_{\rm NCE}$ is the log-probability of the class posterior in a two-class mixture model with a class prior N to M and class distributions p and n. Note that $\hat{c}_{\rm NCE} \neq -\log Z(\hat{\theta}_{\rm NCE})$ and so the model $p(x \mid \hat{\theta}_{\rm NCE}, \hat{c}_{\rm NCE})$ estimated by NCE is not exactly normalized for a finite sample. NCE has consistency and asymptotic normality under mild regularity conditions (Gutmann and Hyvärinen, 2012; Riou-Durand and Chopin, 2018). Note that an idea similar to NCE has been employed in the context of biased sampling (Qin, 2001).

2.2 Bregman divergence related to NCE

Here, we explain the observation by Gutmann and Hirayama (2011) that NCE is interpreted as a projection with respect to a Bregman divergence.

We first review the relationship between the maximum likelihood estimator (MLE) and the Kullback–Leibler divergence. For two probability distributions $q(x)$ and $p(x)$, the Kullback–Leibler divergence $D_{\rm KL}(q,p)$ and Kullback–Leibler discrepancy $d_{\rm KL}(q,p)$ from $q(x)$ to $p(x)$ are defined as

$$D_{\rm KL}(q,p) = \int q(x) \log \frac{q(x)}{p(x)} {\rm d}x, \quad d_{\rm KL}(q,p) = -\int q(x) \log p(x) {\rm d}x,$$

respectively. Note that

$$D_{\rm KL}(q,p) = \int q(x) \log q(x) {\rm d}x + d_{\rm KL}(q,p).$$

For $x^{(1)}, \dots, x^{(N)} \sim p(x \mid \theta)$, the MLE is defined as

$$\hat{\theta}_{\rm MLE} = \arg\max_{\theta} \sum_{t=1}^{N} \log p(x_t \mid \theta).$$

Let $\hat{q}(x)$ be the empirical distribution of $x^{(1)}, \dots, x^{(N)}$ and denote $p(x \mid \theta)$ by p_θ. Then,

$$d_{\rm KL}(\hat{q}, p_\theta) = -\frac{1}{N} \sum_{t=1}^{N} \log p(x_t \mid \theta).$$

Therefore, the MLE minimizes the Kullback–Leibler discrepancy between the empirical distribution and the model:

$$\hat{\theta}_{\rm MLE} = \arg\min_{\theta} d_{\rm KL}(\hat{q}, p_\theta).$$

In this sense, the MLE is interpreted as a projection with respect to the Kullback–Leibler divergence.

Now, we present the analogous result for NCE, which is a special case of the general discussion by Gutmann and Hirayama (2011). Consider a Bregman divergence between two nonnegative measures q and p defined as

$$D_{\rm NCE}(q,p) = \int d_f \left(\frac{q(x)}{n(x)}, \frac{p(x)}{n(x)} \right) n(x) {\rm d}x,$$

where $n(x)$ is a probability density, $d_f(a,b) = f(a) - f(b) - f'(b)(a-b)$ and

$$f(x) = x \log x - \left(\frac{M}{N} + x \right) \log \left(1 + \frac{N}{M} x \right). \tag{5}$$

This divergence is decomposed as

$$D_{\mathrm{NCE}}(q,p) = g(q) + d_{\mathrm{NCE}}(q,p),$$

where $g(q)$ is a quantity depending only on q and

$$d_{\mathrm{NCE}}(q,p) = -\int q(x) \log \frac{Np(x)}{Np(x)+Mn(x)} \mathrm{d}x - \frac{M}{N} \int n(y) \log \frac{Mn(y)}{Np(y)+Mn(y)} \mathrm{d}y. \tag{6}$$

Note that $d_{\mathrm{NCE}}(q,p) = 0$ if and only if $q = p$ since f is strictly convex. Then, the objective function $\hat{d}_{\mathrm{NCE}}(\theta, c)$ of NCE in (4) satisfies

$$\mathrm{E}_y \{ \hat{d}_{\mathrm{NCE}}(\theta, c) \} = d_{\mathrm{NCE}}(\hat{q}, p_{\theta,c}),$$

where $\hat{q}$ is the empirical distribution of $x^{(1)}, \ldots, x^{(N)}$, $p_{\theta,c} = p(\cdot \mid \theta, c)$, and E_y denotes the expectation with respect to noise samples $y^{(1)}, \ldots, y^{(M)}$. Thus, NCE is interpreted as minimizing the discrepancy $d_{\mathrm{NCE}}(\hat{q}, p_{\theta,c})$ between the empirical distribution $\hat{q}(x)$ and the model distribution $p(x \mid \theta, c)$. Although we can adopt f other than (5), Uehara et al. (2018) showed that (5) minimizes the asymptotic variance of the estimator among the class of twice continuously differentiable convex functions.

3. Score matching

In this section, we briefly review the score matching estimator (Hyvärinen, 2005), which is a computationally efficient estimation method for non-normalized models of continuous data.

The score matching method is based on a divergence called the Fisher divergence (Lyu, 2009; Gutmann and Hirayama, 2011). For two probability distributions q and p on $\mathbb{R}^d$, the Fisher divergence is defined as

$$D_{\mathrm{F}}(q,p) = \int \sum_{i=1}^{d} \left\{ \frac{\partial}{\partial x_i} \log q(x) - \frac{\partial}{\partial x_i} \log p(x) \right\}^2 q(x) \mathrm{d}x.$$

By using integration by parts, it is transformed as $D_{\mathrm{F}}(q,p) = g(q) + d_{\mathrm{SM}}(q,p)$, where $g(q)$ is a quantity depending only on q and

$$d_{\mathrm{SM}}(q,p) = \int \left[2 \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} \log p(x) + \sum_{i=1}^{d} \left\{ \frac{\partial}{\partial x_i} \log p(x) \right\}^2 \right] q(x) \mathrm{d}x. \tag{7}$$

Now, suppose we have N i.i.d. samples $x^{(1)}, \ldots, x^{(N)}$ from an unknown distribution $q(x)$ and fit the non-normalized model (1). Then, an unbiased estimator of $d_{\mathrm{SM}}(q, p_\theta)$ in (7) is obtained as

$$\hat{d}_{\mathrm{SM}}(\theta) = \frac{1}{N} \sum_{t=1}^{N} \rho_{\mathrm{SM}}(x^{(t)}, \theta),$$

where

$$\rho_{\mathrm{SM}}(x,\theta) = 2\sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} \log \widetilde{p}(x \mid \theta) + \sum_{i=1}^{d} \left\{ \frac{\partial}{\partial x_i} \log \widetilde{p}(x \mid \theta) \right\}^2 .$$

Importantly, we do not need $Z(\theta)$ for computing $\hat{d}_{\mathrm{SM}}(\theta)$. Thus, the score matching estimator is defined as

$$\hat{\theta}_{\mathrm{SM}} = \arg\min_{\theta} \hat{d}_{\mathrm{SM}}(\theta).$$

This estimator has consistency and asymptotic normality under mild regularity conditions (Hyvärinen, 2005).

Hyvärinen (2007) extended score matching to non-normalized models on $\mathbb{R}_+^d = [0,\infty)^d$ by considering the divergence

$$D_{\mathrm{F+}}(q,p) = \int_{\mathbb{R}_+^d} \sum_{i=1}^{d} \left\{ x_i \frac{\partial}{\partial x_i} \log q(x) - x_i \frac{\partial}{\partial x_i} \log p(x) \right\}^2 q(x) \mathrm{d}x.$$

Through a similar argument to the original score matching, the score matching estimator for non-negative data is defined as

$$\hat{\theta}_{\mathrm{SM+}} = \arg\min_{\theta} \hat{d}_{\mathrm{SM+}}(\theta),$$

where

$$\hat{d}_{\mathrm{SM+}}(\theta) = \frac{1}{N} \sum_{t=1}^{N} \rho_{\mathrm{SM+}}(x^{(t)}, \theta),$$

$$\rho_{\mathrm{SM+}}(x,\theta) = \sum_{i=1}^{d} \left[2x_i \frac{\partial}{\partial x_i} \log \widetilde{p}(x \mid \theta) + x_i^2 \frac{\partial^2}{\partial x_i^2} \log \widetilde{p}(x \mid \theta) + x_i^2 \left\{ \frac{\partial}{\partial x_i} \log \widetilde{p}(x \mid \theta) \right\}^2 \right].$$

See Yu et al. (2018, 2019) for recent developments of non-negative score matching.

For exponential families, the objective functions of the score matching estimators reduce to quadratic forms (Hyvärinen, 2007; Forbes and Lauritzen, 2015). Specifically, for an exponential family

$$p(x \mid \theta) = h(x) \exp\left\{ \sum_{k=1}^{m} \theta_k T_k(x) - \psi(\theta) \right\}$$

on $\mathbb{R}^d$ or $\mathbb{R}_+^d$, the function $\rho_{\mathrm{SM}}(x,\theta)$ or $\rho_{\mathrm{SM+}}(x,\theta)$ is given by a quadratic form

$$\frac{1}{2}\theta^\top \Gamma(x) \theta + g(x)^\top \theta + c(x). \tag{8}$$

For the exact forms of $\Gamma(x)$, $g(x)$ and $c(x)$, see Lin et al. (2016)[2]. Thus, the score matching estimator is obtained by solving the linear equation $\left\{ \sum_{t=1}^{N} \Gamma(x^{(t)}) \right\} \hat{\theta} + \sum_{t=1}^{N} g(x^{(t)}) = 0$.

[2] Note that x_{ij}^2 is missing in the first term of (2.15) in Lin et al. (2016).

4. Akaike information criterion (AIC)

In this section, we briefly review the theory of Akaike information criterion. For more details, see Burnham and Anderson (2002) and Konishi and Kitagawa (2008).

Suppose we have N independent and identically distributed (i.i.d.) samples $x^N = (x^{(1)}, \ldots, x^{(N)})$ from an unknown distribution $q(x)$. Based on them, we predict the future observation z from $q(z)$ by using a predictive distribution. For this aim, we assume a parametric distribution $p(x \mid \theta)$ with an unknown parameter $\theta \in \mathbb{R}^k$ and construct a predictive distribution $p(z \mid \hat{\theta}_{\mathrm{MLE}}(x^N))$, where $\hat{\theta}_{\mathrm{MLE}}(x^N)$ is the maximum likelihood estimate of θ from x^N. Then, the distance between the true distribution $q(z)$ and the predictive distribution $p(z \mid \hat{\theta}_{\mathrm{MLE}}(x^N))$ is evaluated by the Kullback–Leibler divergence

$$D_{\mathrm{KL}}\{q, \hat{\theta}_{\mathrm{MLE}}(x^N)\} = \int q(z) \log \frac{q(z)}{p\{z \mid \hat{\theta}_{\mathrm{MLE}}(x^N)\}} \mathrm{d}z.$$

The Kullback–Leibler divergence is decomposed as

$$D_{\mathrm{KL}}\{q, \hat{\theta}_{\mathrm{MLE}}(x^N)\} = \mathrm{E}_z\{\log q(z)\} + d_{\mathrm{KL}}\{q, \hat{\theta}_{\mathrm{MLE}}(x^N)\}, \tag{9}$$

where E_z denotes the expectation with respect to $z \sim q(z)$ and $d_{\mathrm{KL}}\{q, \hat{\theta}_{\mathrm{MLE}}(x^N)\} = -\mathrm{E}_z[\log p\{z \mid \hat{\theta}_{\mathrm{MLE}}(x^N)\}]$ is the Kullback–Leibler discrepancy from the true distribution $q(z)$ to the predictive distribution $p\{z \mid \hat{\theta}_{\mathrm{MLE}}(x^N)\}$. Since the first term $\mathrm{E}_z\{\log q(z)\}$ in (9) does not depend on $\hat{\theta}_{\mathrm{MLE}}(x^N)$, information criteria are developed as approximately unbiased estimators of the expected Kullback–Leibler discrepancy $\mathrm{E}_x[d_{\mathrm{KL}}\{q, \hat{\theta}_{\mathrm{MLE}}(x^N)\}]$, where E_x denotes the expectation with respect to $x^{(1)}, \ldots, x^{(N)} \sim q(x)$.

Let $\hat{q}$ be the empirical distribution of $x^{(1)}, \ldots, x^{(N)}$. Then, the quantity

$$d_{\mathrm{KL}}\{\hat{q}, \hat{\theta}_{\mathrm{MLE}}(x^N)\} = -\frac{1}{N} \sum_{t=1}^{N} \log p\{x^{(t)} \mid \hat{\theta}_{\mathrm{MLE}}(x^N)\} \tag{10}$$

can be considered as an estimator of $\mathrm{E}_x[d_{\mathrm{KL}}\{q, \hat{\theta}_{\mathrm{MLE}}(x^N)\}]$. However, this simple estimator has negative bias, because the maximum likelihood estimate $\hat{\theta}_{\mathrm{MLE}}(x^N)$ is defined to minimize $d_{\mathrm{KL}}(\hat{q}, \theta)$:

$$\hat{\theta}_{\mathrm{MLE}}(x^N) = \arg\min_{\theta} d_{\mathrm{KL}}(\hat{q}, \theta).$$

If we can compute this negative bias and remove it, then we can actually obtain an unbiased estimator of the Kullback–Leibler discrepancy. This is how typical information criteria are constructed, correcting the inherent bias in using MLE for estimating the Kullback–Leibler discrepancy.

Let $\theta^* = \arg\min_{\theta} d_{\mathrm{KL}}(q, \theta)$. By putting $D_1 = d_{\mathrm{KL}}\{\hat{q}, \hat{\theta}_{\mathrm{MLE}}(x^N)\} - d_{\mathrm{KL}}\{\hat{q}, \theta^*\}$, $D_2 = d_{\mathrm{KL}}\{\hat{q}, \theta^*\} - d_{\mathrm{KL}}\{q, \theta^*\}$ and $D_3 = d_{\mathrm{KL}}\{q, \theta^*\} - d_{\mathrm{KL}}\{q, \hat{\theta}_{\mathrm{MLE}}(x^N)\}$, we have

$$d_{\mathrm{KL}}\{\hat{q}, \hat{\theta}_{\mathrm{MLE}}(x^N)\} - d_{\mathrm{KL}}\{q, \hat{\theta}_{\mathrm{MLE}}(x^N)\} = D_1 + D_2 + D_3. \tag{11}$$

By definition, $\mathrm{E}_x(D_2) = 0$. Also, as $N \to \infty$,

$$N D_1 \xrightarrow{d} -\frac{1}{2} s_1^{\top} J(\theta^*) s_1, \quad N D_3 \xrightarrow{d} -\frac{1}{2} s_3^{\top} J(\theta^*) s_3, \tag{12}$$

where $s_1 \sim \mathrm{N}\left\{0, J(\theta^*)^{-1} I(\theta^*) J(\theta^*)^{-1}\right\}$, $s_3 \sim \mathrm{N}\left\{0, J(\theta^*)^{-1} I(\theta^*) J(\theta^*)^{-1}\right\}$, $k \times k$ matrices $I(\theta)$ and $J(\theta)$ are defined as

$$\begin{aligned} I_{ij}(\theta) &= \mathrm{E}_z \left\{ \frac{\partial}{\partial \theta_i} \log p(z \mid \theta) \frac{\partial}{\partial \theta_j} \log p(z \mid \theta) \right\}, \\ J_{ij}(\theta) &= -\mathrm{E}_z \left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(z \mid \theta) \right\}, \end{aligned}$$

and $J(\theta^*)$ is assumed to be positive definite. Note that the expectation of the limit distribution of ND_1 and ND_3 is

$$-\frac{1}{2} \mathrm{E}\{s_1^\top J(\theta^*) s_1\} = -\frac{1}{2} \mathrm{tr}\left\{ I(\theta^*) J(\theta^*)^{-1} \right\}. \tag{13}$$

From (10), (11), (12) and (13), Takeuchi (1976) proposed

$$\mathrm{TIC} = -2 \sum_{t=1}^{N} \log p\{x^{(t)} \mid \hat{\theta}_{\mathrm{MLE}}(x^N)\} + 2\mathrm{tr}(\hat{I}\hat{J}^{-1}) \tag{14}$$

as an approximately unbiased estimator of $2N\mathrm{E}_x[d_{\mathrm{KL}}\{q, \hat{\theta}_{\mathrm{MLE}}(x^N)\}]$, where $\hat{I}$ and $\hat{J}$ are consistent estimators of $I(\theta^*)$ and $J(\theta^*)$ given by

$$\hat{I}_{ij} = \frac{1}{N} \sum_{t=1}^{N} \frac{\partial}{\partial \theta_i} \log p(x^{(t)} \mid \theta) \frac{\partial}{\partial \theta_j} \log p(x^{(t)} \mid \theta) \bigg|_{\theta = \hat{\theta}_{\mathrm{MLE}}(x^N)},$$

$$\hat{J}_{ij} = -\frac{1}{N} \sum_{t=1}^{N} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(x^{(t)} \mid \theta) \bigg|_{\theta = \hat{\theta}_{\mathrm{MLE}}(x^N)}.$$

The quantity (14) is called Takeuchi Information Criterion.

If the model includes the true distribution: $q(x) = p(x \mid \theta^*)$ for some θ^*, then $I(\theta^*)$ and $J(\theta^*)$ coincide and thus $\mathrm{tr}\{I(\theta^*) J(\theta^*)^{-1}\} = k$. Recall that k is the dimension of the parameter $\theta \in \mathbb{R}^k$. Based on this, Akaike (1974) proposed

$$\mathrm{AIC} = -2 \sum_{t=1}^{N} \log p\{x^{(t)} \mid \hat{\theta}_{\mathrm{MLE}}(x^N)\} + 2k \tag{15}$$

as an approximately unbiased estimator of $2N\mathrm{E}_x[d_{\mathrm{KL}}\{q, \hat{\theta}_{\mathrm{MLE}}(x^N)\}]$. The quantity (15) is called Akaike Information Criterion.

Thus, information criteria enable to compare the goodness of fit of statistical models. Among several candidate models, the model with minimum information criterion is considered to be the closest to the true data-generating process. In practice, since TIC requires more computation than AIC and, furthermore, TIC often suffers from instability caused by estimation errors in $\hat{I}$ and $\hat{J}$, AIC is recommended to use as long as the model is not badly mis-specified (see Burnham and Anderson, 2002, Section 2.3).

5. Information criteria for NCE (NCIC)

In this section, we develop new information criteria for NCE, which we call the Noise Contrastive Information Criterion (NCIC).

5.1 Setting and assumptions

Suppose we have N i.i.d. samples $x^{(1)}, \dots, x^{(N)}$ from an unknown distribution $q(x)$ and estimate a non-normalized model (2) by using NCE with M noise samples $y^{(1)}, \dots, y^{(M)}$ from $n(y)$. The true distribution $q(x)$ may not be contained in the assumed non-normalized model. The parameter θ in the non-normalized model (2) is assumed to be identifiable and have a compact parameter space $\Theta \subset \mathbb{R}^{m-1}$.

For convenience, we denote $\xi = (\theta, c)$, $m = \dim(\xi) = \dim(\theta) + 1$, $\hat{\xi} = \hat{\xi}_{\mathrm{NCE}}$ and $\hat{p}(x) = p(x \mid \hat{\xi})$. Namely, $\hat{\xi} = \hat{\xi}_{\mathrm{NCE}}$ is the minimizer of $\hat{d}_{\mathrm{NCE}}$ in (4). The gradient and Hessian with respect to ξ are written as ∇_ξ and ∇_ξ^2, respectively. Also, we define $\xi^* = \arg\min_\xi d_{\mathrm{NCE}}(q, p_\xi) = (\theta^*, c^*)$ and write $p_*(x) = p(x \mid \xi^*)$. Note that $p_*(x) = q(x)$ when the model includes the true distribution. We denote the expectation with respect to $x^{(1)}, \dots, x^{(N)} \sim q(x)$ and $y^{(1)}, \dots, y^{(M)} \sim n(y)$ by $\mathrm{E}_{x,y}$. The expectation and covariance matrix with respect to $z \sim p(z)$ are denoted by E_p and Cov_p, respectively.

Following Gutmann and Hyvärinen (2012) and Riou-Durand and Chopin (2018), we consider the asymptotics where $N \to \infty$, $M \to \infty$ and $M/N \to \nu$ with $0 < \nu < \infty$. Let

$$\rho_d(x, \xi) = -\log \frac{Np(x \mid \xi)}{Np(x \mid \xi) + Mn(x)}, \tag{16}$$

$$\rho_n(y, \xi) = -\log \frac{Mn(y)}{Np(y \mid \xi) + Mn(y)}. \tag{17}$$

Then, the objective function to be minimized in NCE is represented as

$$\hat{d}_{\mathrm{NCE}}(\xi) = \frac{1}{N} \sum_{t=1}^{N} \rho_d(x^{(t)}, \xi) + \frac{1}{N} \sum_{t=1}^{M} \rho_n(y^{(t)}, \xi).$$

Define $m \times m$ matrices $I(\xi)$ and $J(\xi)$ by

$$I(\xi) = \frac{N}{N+M} \mathrm{Cov}_q \left\{ \nabla_\xi \rho_d(z, \xi) \right\} + \frac{M}{N+M} \mathrm{Cov}_n \left\{ \nabla_\xi \rho_n(z, \xi) \right\},$$

$$J(\xi) = \frac{N}{N+M} \mathrm{E}_q \left\{ \nabla_\xi^2 \rho_d(z, \xi) \right\} + \frac{M}{N+M} \mathrm{E}_n \left\{ \nabla_\xi^2 \rho_n(z, \xi) \right\}.$$

We assume the following regularity conditions:

(N1) For every θ, the support of $p(z \mid \theta)$ is included in that of $n(z)$.

(N2) For every z, $\log p(z \mid \theta)$ is three times continuously differentiable over Θ.

(N3) Both $\mathrm{E}_q \left\{ \nabla_\xi \rho_d(z, \xi^*) \nabla_\xi \rho_d(z, \xi^*)^\top \right\}$ and $\mathrm{E}_n \left\{ \nabla_\xi \rho_n(z, \xi^*) \nabla_\xi \rho_n(z, \xi^*)^\top \right\}$ are finite.

(N4) There exist functions $b_d(z)$ and $b_n(z)$ such that

$$\begin{aligned} |\rho_d(z,\xi)| \leq b_d(z), \quad \left|\frac{\partial^2}{\partial\xi_i\partial\xi_j}\rho_d(z,\xi)\right| \leq b_d(z), \quad \left|\frac{\partial^3}{\partial\xi_i\partial\xi_j\partial\xi_k}\rho_d(z,\xi)\right| \leq b_d(z), \\ |\rho_n(z,\xi)| \leq b_n(z), \quad \left|\frac{\partial^2}{\partial\xi_i\partial\xi_j}\rho_n(z,\xi)\right| \leq b_n(z), \quad \left|\frac{\partial^3}{\partial\xi_i\partial\xi_j\partial\xi_k}\rho_n(z,\xi)\right| \leq b_n(z) \end{aligned}$$

for all i, j, k and ξ, where $\mathrm{E}_q\{b_d(z)\} < \infty$ and $\mathrm{E}_n\{b_n(z)\} < \infty$.

(N5) The matrix $J(\xi^*)$ is nonsingular.

Assumption (N1) is standard in NCE (Gutmann and Hyvärinen, 2012). Assumptions (N2)-(N5) are similar to the regularity conditions for AIC and TIC (Konishi and Kitagawa, 2008). Note that $d_{\mathrm{NCE}}(q, p_*)$ is finite from the assumption (N4).

5.2 Bias evaluation

Similarly to $d_{\mathrm{KL}}\{\hat{q}, \hat{\theta}_{\mathrm{MLE}}(x^N)\}$ in (10), the quantity $\hat{d}_{\mathrm{NCE}}(\hat{\xi})$ has negative bias as an estimator of $\mathrm{E}_{x,y}\{d_{\mathrm{NCE}}(q, \hat{p})\}$. Here, we evaluate this bias following a similar argument to AIC and TIC (Burnham and Anderson, 2002; Konishi and Kitagawa, 2008).

First, the asymptotic distribution of NCE is obtained as follows.

Lemma 1 *Under (N1)-(N5),*

$$\sqrt{N}\left(\hat{\xi} - \xi^*\right) \xrightarrow{d} \mathrm{N}\left\{0, J(\xi^*)^{-1} I(\xi^*) J(\xi^*)^{-1}\right\}. \tag{18}$$

Proof The current setting of NCE corresponds to stratified sampling with two strata: data (size N) and noise (size M). From Theorem 3.1 of Wooldridge (2001), $\hat{\xi}$ is consistent: $\hat{\xi} \xrightarrow{p} \xi^*$. Then, from Theorem 3.2 of Wooldridge (2001), the asymptotic distribution of $\hat{\xi}$ is obtained as (18). ■

Note that (18) is valid even when the model is mis-specified. We also note that Riou-Durand and Chopin (2018) established a rigorous asymptotic theory of NCE under general MCMC sampling of noise.

Let $D_1 = \hat{d}_{\mathrm{NCE}}(\hat{\xi}) - \hat{d}_{\mathrm{NCE}}(\xi^*)$, $D_2 = \hat{d}_{\mathrm{NCE}}(\xi^*) - d_{\mathrm{NCE}}(q, p_*)$ and $D_3 = d_{\mathrm{NCE}}(q, p_*) - d_{\mathrm{NCE}}(q, \hat{p})$. Then,

$$\hat{d}_{\mathrm{NCE}}(\hat{\xi}) - d_{\mathrm{NCE}}(q, \hat{p}) = D_1 + D_2 + D_3. \tag{19}$$

Lemma 2 *Under (N1)-(N5),*

(a) $ND_1 \xrightarrow{d} -\frac{1}{2}s^\top J(\xi^*)s$, *where* $s \sim \mathrm{N}\left\{0, J(\xi^*)^{-1} I(\xi^*) J(\xi^*)^{-1}\right\}$.

(b) $\mathrm{E}_{x,y}(D_2) = 0$.

(c) $ND_3 \xrightarrow{d} -\frac{1}{2}s^\top J(\xi^*)s$, *where* $s \sim \mathrm{N}\left\{0, J(\xi^*)^{-1} I(\xi^*) J(\xi^*)^{-1}\right\}$.

Proof (a) Since $\nabla_\xi \hat{d}_{\mathrm{NCE}}(\xi) = 0$ at $\xi = \hat{\xi}$, the Taylor expansion of $\hat{d}_{\mathrm{NCE}}(\xi)$ around $\xi = \hat{\xi}$ is given by

$$\hat{d}_{\mathrm{NCE}}(\xi^*) = \hat{d}_{\mathrm{NCE}}(\hat{\xi}) + \frac{1}{2}(\xi^* - \hat{\xi})^\top \nabla_\xi^2 \hat{d}_{\mathrm{NCE}}(\xi^\dagger)(\xi^* - \hat{\xi}),$$

where $\xi^\dagger$ is a vector on the segment from $\hat{\xi}$ to ξ^*. From $\hat{\xi} \xrightarrow{p} \xi^*$ and the discussion in Section 3.3 of Wooldridge (2001), $\nabla_\xi^2 \hat{d}_{\mathrm{NCE}}(\xi^\dagger) \xrightarrow{p} J(\xi^*)$. Therefore, from Lemma 1 and Slutsky's theorem,

$$ND_1 = -\frac{N}{2}(\hat{\xi} - \xi^*)^\top \nabla_\xi^2 \hat{d}_{\mathrm{NCE}}(\xi^\dagger)(\hat{\xi} - \xi^*) \xrightarrow{d} -\frac{1}{2} s^\top J(\xi^*) s,$$

where $s \sim \mathrm{N}\left\{0, J(\xi^*)^{-1} I(\xi^*) J(\xi^*)^{-1}\right\}$.

(b) From (4), (6) and the law of large numbers, $\mathrm{E}_{x,y}(D_2) = 0$.

(c) Since $\nabla_\xi d_{\mathrm{NCE}}(q, p_\xi) = 0$ at $\xi = \xi^*$ and $\nabla_\xi^2 d_{\mathrm{NCE}}(q, p_\xi) = J(\xi)$, the Taylor expansion of $d_{\mathrm{NCE}}(q, p_\xi)$ around $\xi = \xi^*$ is given by

$$d_{\mathrm{NCE}}(q, \hat{p}) = d_{\mathrm{NCE}}(q, p_*) + \frac{1}{2}(\hat{\xi} - \xi^*)^\top J(\xi^\dagger)(\hat{\xi} - \xi^*),$$

where $\xi^\dagger$ is a vector on the segment from $\hat{\xi}$ to ξ^*. From $\hat{\xi} \xrightarrow{p} \xi^*$ and the continuous mapping theorem, we have $J(\xi^\dagger) = J(\xi^*) + o_p(1)$. Therefore, from Lemma 1 and Slutsky's theorem,

$$ND_3 = -\frac{N}{2}(\hat{\xi} - \xi^*)^\top J(\xi^\dagger)(\hat{\xi} - \xi^*) \xrightarrow{d} -\frac{1}{2} s^\top J(\xi^*) s,$$

where $s \sim \mathrm{N}\left\{0, J(\xi^*)^{-1} I(\xi^*) J(\xi^*)^{-1}\right\}$. ■

From Lemma 2, the expectation of the limit distribution of ND_1 and ND_3 is

$$-\frac{1}{2}\mathrm{E}\{s^\top J(\xi^*) s\} = -\frac{1}{2}\mathrm{tr}\left\{I(\xi^*) J(\xi^*)^{-1}\right\}. \tag{20}$$

When the model includes the true distribution (well-specified case), (20) has a simpler form. Let

$$b(z) = \frac{p_*(z) n(z)}{r(z)^2},$$

where

$$r(z) = \frac{N}{N+M} p_*(z) + \frac{M}{N+M} n(z). \tag{21}$$

is a mixture distribution of p_* and n.

Lemma 3 *Assume that the model includes the true distribution: $q(x) = p(x \mid \xi^*)$. Then,*

$$\mathrm{tr}\left\{I(\xi^*) J(\xi^*)^{-1}\right\} = m - \mathrm{E}_r\left\{b(z)\right\}, \tag{22}$$

where E_r denotes the expectation with respect to $z \sim r(z)$ in (21).

Proof Let $s(z \mid \xi) = \nabla_\xi \log p(z \mid \xi)$ and $H(z \mid \xi) = \nabla_\xi^2 \log p(z \mid \xi)$. Since $\log p(z \mid \xi) = \log \widetilde{p}(z \mid \theta) + c$ where $\xi = (\theta, c)$, we have $s_m(z \mid \xi) = 1$ and $H_{im}(z \mid \xi) = H_{mi}(z \mid \xi) = 0$ for $i = 1, \dots, m$.

From the definition of ρ_d and ρ_n in (16) and (17),

$$\nabla_\xi \rho_d(z, \xi) = -\frac{Mn(z)}{Np(z \mid \xi) + Mn(z)} s(z \mid \xi), \quad \nabla_\xi \rho_n(z, \xi) = \frac{Np(z \mid \xi)}{Np(z \mid \xi) + Mn(z)} s(z \mid \xi).$$

Thus,

$$\begin{aligned}
\nabla_\xi^2 \rho_d(z, \xi) &= \frac{Np(z \mid \xi) \cdot Mn(z)}{(Np(z \mid \xi) + Mn(z))^2} s(z \mid \xi) s(z \mid \xi)^\top - \frac{Mn(z)}{Np(z \mid \xi) + Mn(z)} H(z \mid \xi), \\
\nabla_\xi^2 \rho_n(z, \xi) &= \frac{Np(z \mid \xi) \cdot Mn(z)}{(Np(z \mid \xi) + Mn(z))^2} s(z \mid \xi) s(z \mid \xi)^\top + \frac{Np(z \mid \xi)}{Np(z \mid \xi) + Mn(z)} H(z \mid \xi).
\end{aligned}$$

Therefore,

$$\begin{aligned}
J(\xi^*) &= \frac{N}{N+M} \int p(z \mid \xi^*) \nabla_\xi^2 \rho_d(z, \xi^*) \mathrm{d}z + \frac{M}{N+M} \int n(z) \nabla_\xi^2 \rho_n(z, \xi^*) \mathrm{d}z \\
&= \frac{1}{N+M} \int \frac{Np(z \mid \xi^*) \cdot Mn(z)}{Np(z \mid \xi^*) + Mn(z)} s(z \mid \xi^*) s(z \mid \xi^*)^\top \mathrm{d}z.
\end{aligned}$$

Since $s_m(z \mid \xi) = 1$, the m-th column vector of $J(\xi^*)$ is

$$j_m(\xi^*) = \frac{1}{N+M} \int \frac{Np(z \mid \xi^*) \cdot Mn(z)}{Np(z \mid \xi^*) + Mn(z)} s(z \mid \xi^*) \mathrm{d}z.$$

Then,

$$\begin{aligned}
I(\xi^*) &= \frac{N}{N+M} \mathrm{Cov}_q \left\{ \nabla_\xi \rho_d(z, \xi^*) \right\} + \frac{M}{N+M} \mathrm{Cov}_n \left\{ \nabla_\xi \rho_n(z, \xi^*) \right\} \\
&= J(\xi^*) - \frac{(N+M)^2}{NM} j_m(\xi^*) j_m(\xi^*)^\top,
\end{aligned}$$

where we used

$$\begin{aligned}
&\mathrm{Cov}_q \left\{ \nabla_\xi \rho_d(z, \xi) \right\} \\
=& \mathrm{E}_q \left\{ \nabla_\xi \rho_d(z, \xi) \nabla_\xi \rho_d(z, \xi)^\top \right\} - \mathrm{E}_q \left\{ \nabla_\xi \rho_d(z, \xi) \right\} \mathrm{E}_q \left\{ \nabla_\xi \rho_d(z, \xi) \right\}^\top \\
=& \int p(z \mid \xi) \left(\frac{Mn(z)}{Np(z \mid \xi) + Mn(z)} \right)^2 s(z \mid \xi) s(z \mid \xi)^\top \mathrm{d}z \\
& \quad - \left(\int p(z \mid \xi) \frac{Mn(z)}{Np(z \mid \xi) + Mn(z)} s(z \mid \xi) \mathrm{d}z \right) \left(\int p(z \mid \xi) \frac{Mn(z)}{Np(z \mid \xi) + Mn(z)} s(z \mid \xi) \mathrm{d}z \right)^\top,
\end{aligned}$$

$$\begin{aligned}
&\mathrm{Cov}_n \left\{ \nabla_\xi \rho_n(z, \xi) \right\} \\
=& \mathrm{E}_n \left\{ \nabla_\xi \rho_n(z, \xi) \nabla_\xi \rho_n(z, \xi)^\top \right\} - \mathrm{E}_n \left\{ \nabla_\xi \rho_n(z, \xi) \right\} \mathrm{E}_q \left\{ \nabla_\xi \rho_n(z, \xi) \right\}^\top \\
=& \int n(z) \left(\frac{Np(z \mid \xi)}{Np(z \mid \xi) + Mn(z)} \right)^2 s(z \mid \xi) s(z \mid \xi)^\top \mathrm{d}z \\
& \quad - \left(\int n(z) \frac{Np(z \mid \xi)}{Np(z \mid \xi) + Mn(z)} s(z \mid \xi) \mathrm{d}z \right) \left(\int n(z) \frac{Np(z \mid \xi)}{Np(z \mid \xi) + Mn(z)} s(z \mid \xi) \mathrm{d}z \right)^\top.
\end{aligned}$$

Thus,

$$\begin{aligned}\operatorname{tr}\left\{I(\xi^*)J(\xi^*)^{-1}\right\} &= m - \frac{(N+M)^2}{NM} j_m(\xi^*)^\top J(\xi^*)^{-1} j_m(\xi^*) \\ &= m - \mathrm{E}_r\left\{b(z)\right\}.\end{aligned}$$

■

Gutmann and Hyvärinen (2012) pointed out that NCE converges to the maximum likelihood estimator as $M/(N+M) \to 1$ and Riou-Durand and Chopin (2018) gave its proof. In this setting, $r(z)$ converges to $n(z)$ and thus $\mathrm{E}_r\left\{b(z)\right\}$ goes to one. As a result, (22) goes to $m-1$, which is equal to the dimension of the parameter θ.

Pan (2001) and Mattheou et al. (2009) proposed information criteria with the quasi-likelihood and density power divergence, respectively, based on similar bias calculations. In comparison, the bias term here takes a more complicated form because we estimate not only the parameter but also the normalization constant in NCE.

5.3 Noise Contrastive Information Criterion (NCIC)

Now, we develop NCIC by using the bias evaluation in the previous subsection.

Let

$$\begin{aligned}\overline{\nabla_\xi \rho_d} &= \frac{1}{N}\sum_{t=1}^{N} \nabla_\xi \rho_d(x^{(t)}, \hat{\xi}), \\ \overline{\nabla_\xi \rho_n} &= \frac{1}{M}\sum_{t=1}^{M} \nabla_\xi \rho_n(y^{(t)}, \hat{\xi}),\end{aligned}$$

and define $m \times m$ matrices $\hat{I}$ and $\hat{J}$ by

$$\begin{aligned}\hat{I} = \frac{1}{N+M}\Bigg[&\sum_{t=1}^{N}\left\{\nabla_\xi \rho_d(x^{(t)}, \hat{\xi}) - \overline{\nabla_\xi \rho_d}\right\}\left\{\nabla_\xi \rho_d(x^{(t)}, \hat{\xi}) - \overline{\nabla_\xi \rho_d}\right\}^\top \\ &+ \sum_{t=1}^{M}\left\{\nabla_\xi \rho_n(y^{(t)}, \hat{\xi}) - \overline{\nabla_\xi \rho_n}\right\}\left\{\nabla_\xi \rho_n(y^{(t)}, \hat{\xi}) - \overline{\nabla_\xi \rho_n}\right\}^\top\Bigg],\end{aligned}$$

$$\hat{J} = \frac{1}{N+M}\left\{\sum_{t=1}^{N} \nabla_\xi^2 \rho_d(x^{(t)}, \hat{\xi}) + \sum_{t=1}^{M} \nabla_\xi^2 \rho_n(y^{(t)}, \hat{\xi})\right\}.$$

From the discussion in Section 3.3 of Wooldridge (2001), $\hat{I}$ and $\hat{J}$ are consistent estimators of $I(\xi^*)$ and $J(\xi^*)$, respectively. Thus, from (19) and Lemma 2, we propose the quantity

$$\mathrm{NCIC}_1 = N\hat{d}_{\mathrm{NCE}}(\hat{\xi}_{\mathrm{NCE}}) + \operatorname{tr}(\hat{I}\hat{J}^{-1}) \tag{23}$$

as an approximately unbiased estimator of $N\mathrm{E}_{x,y}\left\{d_{\mathrm{NCE}}(q, \hat{p})\right\}$.

We also propose a simpler version of NCIC by assuming that the model includes the true distribution. Let

$$\hat{b}(z) = \frac{\hat{p}(z)n(z)}{\hat{r}(z)^2}, \tag{24}$$

where

$$\hat{r}(z) = \frac{N}{N+M}\hat{p}(z) + \frac{M}{N+M}n(z).$$

Then, from Lemma 3, we propose the quantity

$$\mathrm{NCIC}_2 = N\hat{d}_{\mathrm{NCE}}(\hat{\xi}_{\mathrm{NCE}}) + m - \frac{1}{N+M}\left\{\sum_{t=1}^{N}\hat{b}(x^{(t)}) + \sum_{t=1}^{M}\hat{b}(y^{(t)})\right\} \tag{25}$$

as an approximately unbiased estimator of $N\mathrm{E}_{x,y}\{d_{\mathrm{NCE}}(q,\hat{p})\}$.

By minimizing NCIC, we can select from non-normalized models (2) estimated by NCE. NCIC_1 (23) and NCIC_2 (25) are viewed as analogues of TIC (14) and AIC (15) for non-normalized models, respectively. As will be shown in Section 7.1, NCIC_2 has much smaller variance than NCIC_1. Also, NCIC_2 is computationally more efficient than NCIC_1. Therefore, NCIC_2 is recommended to use when the model is considered to be not badly misspecified. This situation is quite similar to that of TIC and AIC (see Burnham and Anderson, 2002, Section 2.3).

Since NCE is an M-estimator, we can also develop Generalized Information Criterion (GIC; Konishi and Kitagawa, 1996) for NCE in principle. However, GIC involves the log-likelihood of the model and thus requires to compute the intractable normalization constant. On the other hand, NCIC can be readily calculated from the result of NCE.

Instead of NCIC, we can also use leave-one-out cross-validation (LOOCV) with NCE for model selection. Specifically[3], for $t = 1, \ldots, N$, let $\hat{\xi}^{(-t)}$ be the estimate of ξ by NCE applied to $x^{(1)}, \ldots, x^{(t-1)}, x^{(t+1)}, \ldots, x^{(N)}$ and $y^{(1)}, \ldots, y^{(t-1)}, y^{(t+1)}, \ldots, y^{(N)}$. Then, the quantity

$$\text{NCE-CV} = \sum_{t=1}^{N} \rho_d(x^{(t)}, \hat{\xi}^{(-t)}) + \sum_{t=1}^{M} \rho_n(y^{(t)}, \hat{\xi}^{(-t)}). \tag{26}$$

can be adopted as an approximately unbiased estimator of $N\mathrm{E}_{x,y}\{d_{\mathrm{NCE}}(q,\hat{p})\}$. We confirmed by simulation that the model selection performances of NCIC and NCE-CV are comparable, whereas NCIC is computationally more efficient than NCE-CV (Section 7.3).

In developing NCIC, we assumed that the noise samples are independent. Recently, Riou-Durand and Chopin (2018) established the asymptotic theory of NCE including cases where the noise samples are generated by MCMC. It is an interesting future work to extend NCIC to such general cases. Further problems for future research include extension to generalized NCE (Uehara et al., 2020a) and missing data (Uehara et al., 2020b).

[3] Here, we assume $M = N$ for convenience.

6. Information criteria for score matching (SMIC)

In this section, we develop new information criteria for score matching, which we call the Score Matching Information Criterion (SMIC). For convenience, we focus on the original score matching estimator $\hat{\theta}_{\mathrm{SM}}$ in the following. Analogous results for the score matching estimator $\hat{\theta}_{\mathrm{SM}+}$ for non-negative data are obtained by replacing $\hat{d}_{\mathrm{SM}}$ and ρ_{SM} with $\hat{d}_{\mathrm{SM}+}$ and $\rho_{\mathrm{SM}+}$, respectively.

Suppose we have N i.i.d. samples $x^{(1)}, \ldots, x^{(N)}$ from an unknown distribution $q(x)$ and fit a non-normalized model (1) with $\theta \in \mathbb{R}^k$ by score matching. Here, the true distribution $q(x)$ may not be contained in the assumed non-normalized model. We define $\theta^* = \arg\min_\theta d_{\mathrm{SM}}(q, p_\theta)$ and write $p_*(x) = p(x \mid \theta^*)$ and $\hat{p}(x) = p(x \mid \hat{\theta}_{\mathrm{SM}})$. Note that $p_*(x) = q(x)$ when the model includes the true distribution.

Define $k \times k$ matrices $I(\theta)$ and $J(\theta)$ by

$$I(\theta) = \mathrm{Cov}_q \left\{ \nabla_\theta \rho_{\mathrm{SM}}(x, \theta) \right\}, \quad J(\theta) = \mathrm{E}_q \left\{ \nabla_\theta^2 \rho_{\mathrm{SM}}(x, \theta) \right\}.$$

Assume the following regularity conditions:

(S1) For every x, $\log p(x \mid \theta)$ is C^3 with respect to θ.

(S2) $\mathrm{E}_q \left[\nabla_\theta \rho_{\mathrm{SM}}(x, \theta^*) \nabla_\theta \rho_{\mathrm{SM}}(x, \theta^*)^\top \right]$ is finite.

(S3) There exists a function $b(x)$ such that

$$|\rho_{\mathrm{SM}}(x, \theta)| \leq b(x), \quad \left| \frac{\partial^2}{\partial \theta_i \partial \theta_j} \rho_{\mathrm{SM}}(x, \theta) \right| \leq b(x), \quad \left| \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} \rho_{\mathrm{SM}}(x, \theta) \right| \leq b(x),$$

for all i, j, k and θ, where $\mathrm{E}_q[b(x)] < \infty$.

(S4) The matrix $J(\theta^*)$ is nonsingular.

The quantity $\hat{d}_{\mathrm{SM}}(\hat{\theta}_{\mathrm{SM}})$ has negative bias as an estimator of $\mathrm{E}_x[d_{\mathrm{SM}}(q, \hat{p})]$ and we evaluate this bias following Section 5.2. Let $D_1 = \hat{d}_{\mathrm{SM}}(\hat{\theta}_{\mathrm{SM}}) - \hat{d}_{\mathrm{SM}}(\theta^*)$, $D_2 = \hat{d}_{\mathrm{SM}}(\theta^*) - d_{\mathrm{SM}}(q, p_*)$ and $D_3 = d_{\mathrm{SM}}(q, p_*) - d_{\mathrm{SM}}(q, \hat{p})$. Then,

$$\hat{d}_{\mathrm{SM}}(\hat{\theta}_{\mathrm{SM}}) - d_{\mathrm{SM}}(q, \hat{p}) = D_1 + D_2 + D_3.$$

By using a similar argument to Lemma 2, we obtain the following.

Lemma 4 *Under (S1)-(S4),*

(a) $N D_1 \xrightarrow{d} -\frac{1}{2} s^\top J(\xi^*) s$, *where* $s \sim \mathrm{N}\left\{0, J(\xi^*)^{-1} I(\xi^*) J(\xi^*)^{-1}\right\}$.

(b) $\mathrm{E}_q(D_2) = 0$.

(c) $N D_3 \xrightarrow{d} -\frac{1}{2} s^\top J(\xi^*) s$, *where* $s \sim \mathrm{N}\left\{0, J(\xi^*)^{-1} I(\xi^*) J(\xi^*)^{-1}\right\}$.

The expectation of the limit distribution of ND_1 and ND_3 is

$$-\frac{1}{2}\mathrm{E}[s^\top J(\theta^*)s] = -\frac{1}{2}\mathrm{tr}\left\{I(\theta^*)J(\theta^*)^{-1}\right\}.$$

Let

$$\hat{I} = \frac{1}{N}\sum_{t=1}^{N} \nabla_\theta \rho_{\mathrm{SM}}(x^{(t)},\theta)\nabla_\theta \rho_{\mathrm{SM}}(x^{(t)},\theta)^\top \bigg|_{\theta=\hat{\theta}}, \quad \hat{J} = \frac{1}{N}\sum_{t=1}^{N} \nabla_\theta^2 \rho_{\mathrm{SM}}(x^{(t)},\theta)\bigg|_{\theta=\hat{\theta}}.$$

Then, $\hat{I}$ and $\hat{J}$ are consistent estimators of $I(\theta^*)$ and $J(\theta^*)$, respectively. Thus, from Lemma 4, we propose the quantity

$$\mathrm{SMIC} = N\hat{d}_{\mathrm{SM}}(\hat{\theta}_{\mathrm{SM}}) + \mathrm{tr}(\hat{I}\hat{J}^{-1}) \tag{27}$$

as an approximately unbiased estimator of $N\mathrm{E}_q\{d_{\mathrm{SM}}(q,\hat{p})\}$.

For exponential families, the function $\rho_{\mathrm{SM}}(x,\theta)$ is given by the quadratic form (8) and so $\hat{I}$ and $\hat{J}$ in (27) become simple:

$$\hat{I} = \frac{1}{N}\sum_{t=1}^{N}\left\{\Gamma(x^{(t)})\hat{\theta} + g(x^{(t)})\right\}\left\{\Gamma(x^{(t)})\hat{\theta} + g(x^{(t)})\right\}^\top, \quad \hat{J} = \frac{1}{N}\sum_{t=1}^{N}\Gamma(x^{(t)}).$$

On the other hand, for general models, the term $\hat{J}$ in SMIC involves fourth-order partial derivatives. Thus, the analytical complexity of SMIC can be larger than NCIC in general. Also note that SMIC can be used only for continuous models whereas NCIC is applicable to both continuous and discrete models.

Unlike Lemma 3 for NCIC, it seems difficult to simplify SMIC in the well-specified case due to the derivative with respect to x in the objective functions of score matching. Also, note that our focus here is different from Dawid and Musio (2015) and Shao et al. (2019), who applied the idea of score matching to Bayesian model selection with improper priors. Finally, whereas we can also develop Generalized Information Criterion (GIC; Konishi and Kitagawa, 1996) for score matching in principle, GIC is based on the log-likelihood of the model and thus requires to compute the intractable normalization constant, which is not necessary in SMIC.

Instead of SMIC, we can, again, use leave-one-out cross-validation (LOOCV) with score matching for model selection. Specifically, for $t = 1,\dots,N$, let $\hat{\theta}^{(-t)}$ be the estimate of θ by score matching applied to $x^{(1)},\dots,x^{(t-1)},x^{(t+1)},\dots,x^{(N)}$. Then, the quantity

$$\text{SM-CV} = \sum_{t=1}^{N}\rho_{\mathrm{SM}}(x^{(t)},\hat{\theta}^{(-t)}). \tag{28}$$

can be adopted as an approximately unbiased estimator of $N\mathrm{E}_q\{d_{\mathrm{SM}}(q,\hat{p})\}$. We confirmed by simulation that the model selection performances of SMIC and SM-CV are comparable, whereas SMIC is computationally more efficient than SM-CV (Section 7.3) similarly to NCIC.

Recently, Liu et al. (2019) developed an estimation method for non-normalized models called the Discriminative Likelihood Estimation (DLE), which approximates the Kullback–Leiber divergence by using the techniques of density ratio estimation and Stein operators. They also proposed an information criterion based on DLE. It is an interesting future work to investigate the relationship of their method with NCIC and SMIC.

7. Simulation results

In this section, we confirm the validity of the proposed information criteria (NCIC_1, NCIC_2, and SMIC) by simulation. For numerical optimization in NCE and score matching, we use the nonlinear conjugate gradient method (Rasmussen, 2006).

7.1 Accuracy of bias correction

First, we check the accuracy of the bias correction terms in NCIC and SMIC.

7.1.1 NCIC

We generated $N = 10^3$ independent samples from the two-component Gaussian mixture distribution $(1-\varepsilon) \cdot \mathrm{N}(0,1) + \varepsilon \cdot \mathrm{N}(0,10)$, where ε specifies the proportion of outliers. Then, we applied NCE to estimate the parameters of the non-normalized model

$$p(x \mid \theta) \propto \exp(\theta_1 x^2 + \theta_2 x), \tag{29}$$

which is a non-normalized version of the Gaussian distribution ($m = 3$). The $M = 10^3$ noise samples were generated from $\mathrm{N}(0,1)$ independently. When $\varepsilon = 0$, the true distribution is included in the model (29). This experimental setting follows Konishi and Kitagawa (1996).

In Section 5, NCIC_1 and NCIC_2 were developed by correcting the bias of the quantity $N\hat{d}_{\mathrm{NCE}}(\hat{\xi}_{\mathrm{NCE}})$ as an estimator of $N\mathrm{E}_{x,y}[d_{\mathrm{NCE}}(q,\hat{p})]$. Namely, the true bias is

$$B = N\mathrm{E}_{x,y}\left\{\hat{d}_{\mathrm{NCE}}(\hat{\xi}_{\mathrm{NCE}})\right\} - N\mathrm{E}_{x,y}\left\{d_{\mathrm{NCE}}(q,\hat{p})\right\},$$

and NCIC_1 in (23) and NCIC_2 in (25) are based on the bias estimates

$$\hat{B}_1 = -\mathrm{tr}(\hat{I}\hat{J}^{-1}),$$

and

$$\hat{B}_2 = -m + \frac{1}{N+M}\left\{\sum_{t=1}^{N} \hat{b}(x^{(t)}) + \sum_{t=1}^{M} \hat{b}(y^{(t)})\right\},$$

respectively. We compare these values numerically by a Monte Carlo simulation with 10^5 repetitions.

Figure 1 plots B, $\mathrm{E}_{x,y}(\hat{B}_1)$ and $\mathrm{E}_{x,y}(\hat{B}_2)$ as a function of ε. When $\varepsilon = 0$ (well-specified case), the bias B is approximately equal to $-(m-1) = -2$ and both $\mathrm{E}_{x,y}(\hat{B}_1)$ and $\mathrm{E}_{x,y}(\hat{B}_2)$ are close to this value. When $\varepsilon > 0$ (mis-specified case), B and $\mathrm{E}_{x,y}(\hat{B}_1)$ coincide quite well. These results are consistent with Lemma 2 and 3. Whereas the standard deviation of $\hat{B}_1$ is around 0.1 (see dotted lines in Figure 1), that of $\hat{B}_2$ is smaller than 10^{-8}. Thus, NCIC_2 has much smaller variance than NCIC_1. This is analogous to the fact that TIC has much larger variance than AIC (Burnham and Anderson, 2002). Interestingly, the absolute bias $|B|$ decreases with ε, whereas it increases with ε for normalized models (see Fig. 1 of Konishi and Kitagawa, 1996). Thus, just using the number of parameters as the bias correction term may be fairly useful and robust in practice, just like AIC is attractive in that the bias correction term is the number of parameters. Therefore, NCIC_2 is recommended to use when the model is considered to be not badly mis-specified. This situation is quite similar to that of TIC and AIC (see Burnham and Anderson, 2002, Section 2.3).

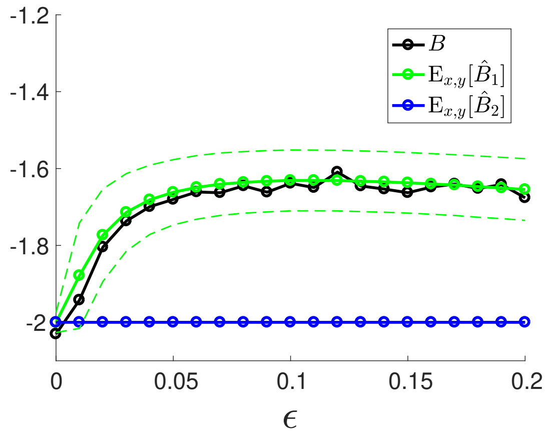

Figure 1: Comparison of the true bias B (black) and the bias estimates $\hat{B}_1$ (green, with standard deviation) and $\hat{B}_2$ (blue, standard deviation $< 10^{-8}$) in NCIC. Here, $\varepsilon = 0$ means a well-specified model.

7.1.2 SMIC

We generated $N = 10^3$ independent samples from the two-component Gaussian mixture distribution $(1-\varepsilon) \cdot \mathrm{N}(0,1) + \varepsilon \cdot \mathrm{N}(0,10)$. Then, we applied score matching to fit the normal distribution (29). When $\varepsilon = 0$, the true distribution is included in the model (29). This experimental setting follows Konishi and Kitagawa (1996). In this case, the model is exponential family and the functions in (8) is

$$\Gamma(x) = \frac{1}{N} \sum_{t=1}^{N} \begin{pmatrix} 8x_t^2 & 4x_t \\ 4x_t & 2 \end{pmatrix}, \quad g(x) = \begin{pmatrix} 4 \\ 0 \end{pmatrix}, \quad c(x) = 0.$$

In SMIC, the true bias $B = N\mathrm{E}_q\{\hat{d}_{\mathrm{SM}}(\hat{\theta}_{\mathrm{SM}})\} - N\mathrm{E}_q\{d_{\mathrm{SM}}(q, \hat{p})\}$ is estimated by $\hat{B} = -\mathrm{tr}(\hat{I}\hat{J}^{-1})$. Figure 2 plots B and $\mathrm{E}_q(\hat{B})$ as a function of ε. These values were computed by a Monte Carlo simulation with 10^5 repetitions. Consistent with Lemma 4, B and $\mathrm{E}_q(\hat{B})$ coincide quite well. Note that the bias goes down before going up again as ϵ increases.

7.2 Gaussian graphical model

Next, we apply NCIC and SMIC to edge selection of the Gaussian graphical model (GGM) (Lauritzen, 1996) and compare their performance with AIC.

Let $G = (V, E)$ be an undirected graph where $V = \{1, \ldots, d\}$. Then, the GGM with graph G is defined as

$$p(x \mid \Sigma) = \frac{1}{(2\pi)^{d/2} (\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2} x^\top \Sigma^{-1} x \right), \quad x \in \mathbb{R}^d, \tag{30}$$

where $\Sigma \in \mathbb{R}^{d \times d}$ is a positive definite matrix satisfying $(\Sigma^{-1})_{ij} = 0$ for $(i,j) \notin E$ and the normalization constant is obtained in closed form. The zero-nonzero pattern of the

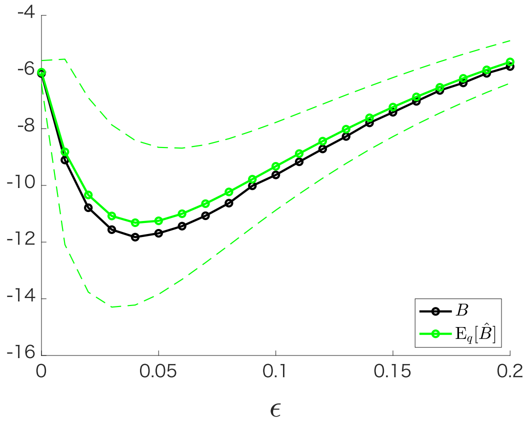

Figure 2: Comparison of the true bias B (black) and the bias estimate $\hat{B}$ (green, with standard deviation) in SMIC.

precision matrix Σ^{-1} specifies the conditional independence structure of $X = (X_1, \cdots, X_d)$: if $(\Sigma^{-1})_{ij} = 0$ $(i \neq j)$, then X_i and X_j are independent conditionally on the other variables X_k $(k \neq i, j)$. Thus, we consider selection of the graph G.

Following Drton and Perlman (2004), we generated N independent samples $x^{(1)}, \ldots, x^{(N)}$ from $\mathrm{N}(0, \Sigma)$ with

$$\Sigma^{-1} = \begin{pmatrix} 1 & \sigma^{12} & 0 \\ \sigma^{12} & 1 & 0.55 \\ 0 & 0.55 & 1 \end{pmatrix}, \tag{31}$$

where the value of σ^{12} is set to 0.2, 0.3 or 0.5. This distribution corresponds to the GGM (30) with the path graph of size $d = 3$: $G_3 = (V_3, E_3)$ where $V_3 = \{1, 2, 3\}$ and $E_3 = \{(1, 2), (2, 3)\}$. Then, we fitted $2^{d(d-1)/2} = 8$ GGMs (30) with each possible G to $x^{(1)}, \ldots, x^{(N)}$ by using NCE, score matching or maximum likelihood estimation (MLE). Namely, we estimated both diagonal and off-diagonal elements of Σ^{-1} under the constraint $(\Sigma^{-1})_{ij} = 0$ for $(i, j) \notin E$. For NCE, we generated $M = N$ noise samples $y^{(1)}, \ldots, y^{(M)}$ from the normal distribution with the same mean and covariance with $x^{(1)}, \ldots, x^{(N)}$. For MLE, we used CVX, a MATLAB package for convex programming (Grant and Boyd, 2018). We selected G that corresponds to the GGM with the minimum NCIC_1, NCIC_2, SMIC or AIC. We repeated the simulation 1000 times.

Tables 1–3 present the detection probabilities of each edge for $N = 100$, $N = 200$ and $N = 1000$, respectively. For all criteria, the edges in G_3, namely $(1, 2)$ and $(2, 3)$, are selected more frequently than the edge absent in G_3, namely $(1, 3)$, especially when N is large. Furthermore, the frequency of selecting the edge $(1, 2)$ increases with the magnitude of σ^{12}. SMIC attains almost the same performance with AIC, which is consistent with the fact that the score matching estimator coincides with the MLE for Gaussian models (Hyvärinen, 2005). On the other hand, the performance of NCIC is a little worse than

Table 1: Detection probabilities of each edge in the GGM (30) for $N = 100$ when (a) $\sigma^{12} = 0.2$, (b) $\sigma^{12} = 0.3$ and (c) $\sigma^{12} = 0.5$. The true edges are $(1, 2)$ and $(2, 3)$.

(a)

	NCIC$_1$	NCIC$_2$	SMIC	AIC
(1,2)	0.515	0.481	0.790	0.783
(1,3)	0.187	0.170	0.199	0.210
(2,3)	0.945	0.928	1.000	1.000

(b)

	NCIC$_1$	NCIC$_2$	SMIC	AIC
(1,2)	0.750	0.706	0.966	0.971
(1,3)	0.198	0.181	0.167	0.165
(2,3)	0.947	0.930	1.000	1.000

(c)

	NCIC$_1$	NCIC$_2$	SMIC	AIC
(1,2)	0.943	0.926	1.000	1.000
(1,3)	0.190	0.171	0.145	0.145
(2,3)	0.953	0.941	1.000	1.000

Table 2: Detection probabilities of each edge in the GGM (30) for $N = 200$ when (a) $\sigma^{12} = 0.2$, (b) $\sigma^{12} = 0.3$ and (c) $\sigma^{12} = 0.5$. The true edges are $(1, 2)$ and $(2, 3)$.

(a)

	NCIC$_1$	NCIC$_2$	SMIC	AIC
(1,2)	0.749	0.719	0.938	0.936
(1,3)	0.218	0.210	0.167	0.170
(2,3)	1.000	0.999	1.000	1.000

(b)

	NCIC$_1$	NCIC$_2$	SMIC	AIC
(1,2)	0.937	0.925	1.000	0.999
(1,3)	0.172	0.162	0.123	0.137
(2,3)	1.000	1.000	1.000	1.000

(c)

	NCIC$_1$	NCIC$_2$	SMIC	AIC
(1,2)	0.997	0.997	1.000	1.000
(1,3)	0.147	0.138	0.147	0.139
(2,3)	0.997	0.997	1.000	1.000

SMIC and AIC, which is reasonable because NCE has larger asymptotic variance than MLE (Uehara et al., 2018).

Table 3: Detection probabilities of each edge in the GGM (30) for $N = 1000$ when (a) $\sigma^{12} = 0.2$, (b) $\sigma^{12} = 0.3$ and (c) $\sigma^{12} = 0.5$. The true edges are $(1, 2)$ and $(2, 3)$.

(a)

	NCIC$_1$	NCIC$_2$	SMIC	AIC
(1,2)	0.999	0.999	1.000	1.000
(1,3)	0.167	0.162	0.145	0.143
(2,3)	1.000	1.000	1.000	1.000

(b)

	NCIC$_1$	NCIC$_2$	SMIC	AIC
(1,2)	1.000	1.000	1.000	1.000
(1,3)	0.169	0.166	0.155	0.159
(2,3)	1.000	1.000	1.000	1.000

(c)

	NCIC$_1$	NCIC$_2$	SMIC	AIC
(1,2)	1.000	1.000	1.000	1.000
(1,3)	0.150	0.148	0.147	0.142
(2,3)	1.000	1.000	1.000	1.000

7.3 Truncated Gaussian graphical model

Now, we apply NCIC and SMIC to edge selection of the truncated Gaussian graphical model (Lin et al., 2016), which has an intractable normalization constant.

For an undirected graph $G = (V, E)$ with $V = \{1, \dots, d\}$, the truncated GGM with graph G is defined as

$$p(x \mid \Sigma) \propto \exp\left(-\frac{1}{2} x^\top \Sigma^{-1} x\right), \quad x \in \mathbb{R}^d_+, \tag{32}$$

where $\Sigma \in \mathbb{R}^{d \times d}$ is a positive definite matrix satisfying $(\Sigma^{-1})_{ij} = 0$ for $(i, j) \notin E$. Due to the truncation to the positive orthant $\mathbb{R}^d_+$, the normalization constant of the truncated GGM (32) is computationally intractable. Similarly to the original GGM (30), X_i and X_j are independent conditionally on the other variables X_k $(k \neq i, j)$ if $(\Sigma^{-1})_{ij} = 0$. Thus, we consider selection of the graph G.

Similarly to the previous subsection, we considered edge selection from N independent samples $x^{(1)}, \dots, x^{(N)}$ from a truncated GGM (32) with covariance (31) where σ^{12} is set to 0.2, 0.3 or 0.5. For NCE, we generated $M = N$ noise samples $y^{(1)}, \dots, y^{(M)}$ from the product of the coordinate-wise exponential distributions with the same mean as $x^{(1)}, \dots, x^{(N)}$. We selected G that corresponds to the truncated GGM with the minimum NCIC$_1$, NCIC$_2$ or SMIC. For comparison with model selection by leave-one-out cross-validation (LOOCV), we also selected G by minimizing NCE-CV (26) or SM-CV (28). We repeated the simulation 1000 times.

Tables 4–6 present the detection probabilities of each edge for $N = 100$, $N = 200$ and $N = 1000$, respectively. The behaviors of NCIC and SMIC are qualitatively the same with Tables 1–3. Namely, the true edges $(1, 2)$ and $(2, 3)$ are selected more frequently

Table 4: Detection probabilities of each edge in the truncated GGM (32) for $N = 100$ when (a) $\sigma^{12} = 0.2$, (b) $\sigma^{12} = 0.3$ and (c) $\sigma^{12} = 0.5$. The true edges are $(1, 2)$ and $(2, 3)$.

(a)

	NCIC$_1$	NCIC$_2$	NCE-CV	SMIC	SM-CV
(1,2)	0.287	0.221	0.210	0.360	0.243
(1,3)	0.187	0.155	0.132	0.303	0.184
(2,3)	0.495	0.429	0.422	0.617	0.513

(b)

	NCIC$_1$	NCIC$_2$	NCE-CV	SMIC	SM-CV
(1,2)	0.316	0.251	0.249	0.449	0.337
(1,3)	0.205	0.182	0.152	0.304	0.205
(2,3)	0.522	0.451	0.449	0.603	0.506

(c)

	NCIC$_1$	NCIC$_2$	NCE-CV	SMIC	SM-CV
(1,2)	0.460	0.405	0.394	0.592	0.478
(1,3)	0.206	0.173	0.147	0.308	0.204
(2,3)	0.496	0.427	0.430	0.594	0.498

than the false edge $(1, 3)$ especially when N is large, and the frequency of selecting the edge $(1, 2)$ increases with the magnitude of σ^{12}. Also, the model selection performances of NCIC and SMIC are comparable to those of NCE-CV and SM-CV, respectively, which is analogous to the asymptotic equivalence of model selection by AIC and LOOCV for normalized models (Stone, 1977). Note that NCE-CV and SM-CV take approximately N times more computational cost than NCIC and SMIC, respectively.

To verify the performance of NCIC and SMIC in higher dimension, we conducted another experiment with $N = 1000$, $d = 16$ and G given by the grid graph in Figure 3. The nonzero off-diagonal entries of Σ^{-1} were all set to 0.5 and the diagonal entries of Σ^{-1} were set to a common value so that the minimum eigenvalue of Σ^{-1} is 0.1. Since the number of possible graph structures is too large in this case, we narrowed down the candidate graphs by using a similar procedure to the graphical LASSO. Specifically, we first applied the NCE and score matching with l_1-regularization on the off-diagonal entries of Σ^{-1}. For optimization, we employed the accelerated proximal gradient algorithm[4]. By changing the value of the regularization parameter, a sequence of candidate graphs was obtained for both NCE and score matching. Then, we fitted the graphical models for candidate graphs without regularization to calculate NCIC and SMIC. Note that such a procedure is also used for LASSO (Belloni and Chernozhukov, 2013). For NCE, we generated $M = N$ noise samples $y^{(1)}, \ldots, y^{(M)}$ from the product of the coordinate-wise exponential distributions with the same mean as $x^{(1)}, \ldots, x^{(N)}$. We repeated the simulation 1000 times. Table 7 presents the true positive rate (the probability of selecting the edges in G) and false positive rate (the probability of selecting the edges not in G). It indicates that both NCIC and SMIC select the edges in G much more frequently than those not in G. The detection performance is

[4] We used the MATLAB program from `https://github.com/bodono/apg`.

Table 5: Detection probabilities of each edge in the truncated GGM (32) for $N = 200$ when (a) $\sigma^{12} = 0.2$, (b) $\sigma^{12} = 0.3$ and (c) $\sigma^{12} = 0.5$. The true edges are $(1, 2)$ and $(2, 3)$.

(a)

	NCIC$_1$	NCIC$_2$	NCE-CV	SMIC	SM-CV
(1,2)	0.302	0.256	0.265	0.383	0.302
(1,3)	0.195	0.178	0.156	0.278	0.198
(2,3)	0.689	0.642	0.651	0.730	0.667

(b)

	NCIC$_1$	NCIC$_2$	NCE-CV	SMIC	SM-CV
(1,2)	0.436	0.371	0.392	0.482	0.424
(1,3)	0.191	0.173	0.156	0.251	0.179
(2,3)	0.704	0.659	0.670	0.744	0.682

(c)

	NCIC$_1$	NCIC$_2$	NCE-CV	SMIC	SM-CV
(1,2)	0.647	0.585	0.613	0.713	0.646
(1,3)	0.201	0.185	0.161	0.284	0.191
(2,3)	0.692	0.639	0.670	0.728	0.665

Table 6: Detection probabilities of each edge in the truncated GGM (32) for $N = 1000$ when (a) $\sigma^{12} = 0.2$, (b) $\sigma^{12} = 0.3$ and (c) $\sigma^{12} = 0.5$. The true edges are $(1, 2)$ and $(2, 3)$.

(a)

	NCIC$_1$	NCIC$_2$	NCE-CV	SMIC	SM-CV
(1,2)	0.613	0.586	0.604	0.623	0.599
(1,3)	0.171	0.162	0.162	0.184	0.156
(2,3)	0.996	0.996	0.996	0.993	0.992

(b)

	NCIC$_1$	NCIC$_2$	NCE-CV	SMIC	SM-CV
(1,2)	0.839	0.820	0.830	0.829	0.809
(1,3)	0.163	0.160	0.154	0.193	0.167
(2,3)	0.996	0.996	0.996	0.995	0.993

(c)

	NCIC$_1$	NCIC$_2$	NCE-CV	SMIC	SM-CV
(1,2)	0.985	0.983	0.982	0.971	0.965
(1,3)	0.183	0.172	0.173	0.195	0.170
(2,3)	0.995	0.995	0.995	0.990	0.988

Table 7: True and false positive rates for the truncated GGM (32) with the grid graph.

	NCIC_1	NCIC_2	SMIC
true positive	0.711	0.662	0.759
false positive	0.191	0.156	0.207

not very strong compared to Table 6, which may be related to the difficulty in selecting an appropriate noise distribution in higher dimensions.

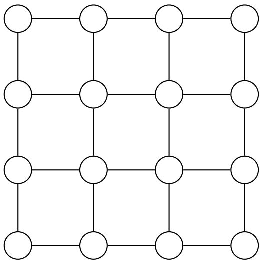

Figure 3: Grid graph.

8. Application to real data

In this section, we apply NCIC and SMIC to model selection for real data of natural image, RNAseq and wind direction.

8.1 Natural image data

First, we apply NCIC to analysis of natural image data with the energy-based overcomplete independent component analysis (ICA) model (Teh et al., 2004) defined by

$$\log p(x \mid w) = \sum_{b=1}^{B} G(w_b^\top x) - \log Z(w_1, \ldots, w_B), \quad x \in \mathbb{R}^d, \tag{33}$$

where $w = (w_1, \ldots, w_B)$ with $w_i \in \mathbb{R}^d$ is the overcomplete set of filters and $G(u) = -|u|$[5]. This model is related to ICA with overcomplete bases (Hyvärinen et al., 2001) and extracts useful features of data. In previous work, the number of filters B ($> d$) has been selected arbitrarily. Here, we determine B by minimizing NCIC.

We used $N = 5 \times 10^4$ image patches of 8×8 pixels taken from natural images. This data is provided in Hoyer's *imageica* package.[6] Following Hyvärinen (2005), we removed

[5] Although this model does not satisfy the smoothness assumption (N2), the non-smoothness here is essentially the same with that in median estimation (van der Vaart, 1998, Example 5.24). The asymptotic variance is still similarly obtained following Theorem 5.23 of van der Vaart (1998) by assuming the smoothness of the expectation of the objective function rather than the objective function itself. We leave the rigorous argument to future work.

[6] http://www.cs.helsinki.fi/patrik.hoyer/

the DC component and then applied whitening. Thus, the data dimension is $d = 63$. For NCE, we used $M = 5 \times 10^4$ noise samples from the Gaussian distribution with the same mean and covariance as data.

Figure 4 (a) plots NCIC$_2$ as a function of B. NCIC$_2$ takes minimum at $B = 118$. Some of the estimated filters $w_1, \dots, w_B$ when $B = 118$ are shown in Figure 4 (b). Here, the filters are converted back to the original space from the whitened space for visualization. Similarly to the result by score matching (Hyvärinen, 2005), many filters represent localized patterns in image patches (Olshausen and Field, 1997; Hyvärinen et al., 2009). Namely, they take (significantly) nonzero value on only limited regions of images. Note that the computation of $\hat{I}$ and $\hat{J}$ in NCIC$_1$ was computationally intractable in this case due to the large sample size N. We did not consider SMIC here because the calculation of $\hat{J}$ in SMIC was analytically complex.

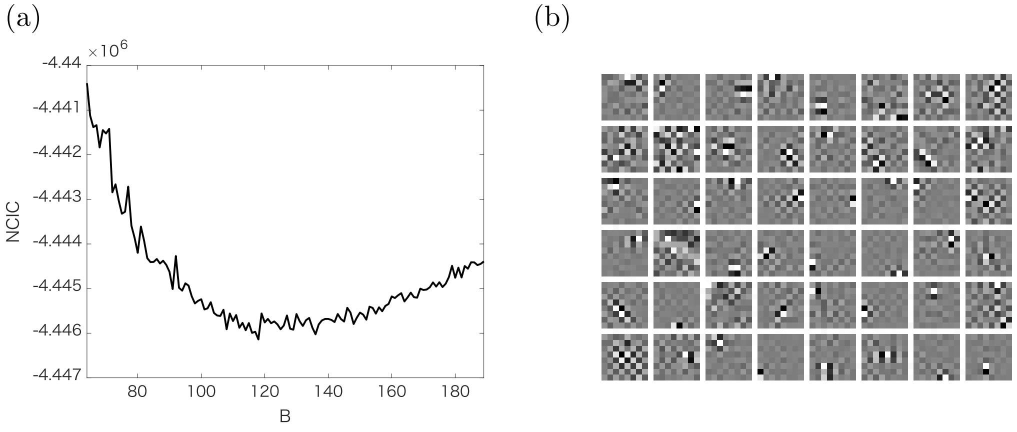

Figure 4: (a) NCIC$_2$ of overcomplete ICA models (33) for natural image data. (b) Estimated filters when $B = 118$.

8.2 RNAseq data

Next, we apply SMIC to comparison of graphical model for the RNAseq data used in Lin et al. (2016). This is a non-negative multivariate data of sample size $N = 487$. We analyze $d = 40$ among 330 genes that do not contain missing values and have coefficient of variation larger than one.

To investigate interaction between genes, Lin et al. (2016) fitted the truncated Gaussian graphical model (32) to RNAseq data by l_1-regularized score matching, which can be solved by the existing algorithms for LASSO. Another possible model is the log-Gaussian graphical model defined by

$$p(x \mid \mu, \Sigma) \propto \left(\prod_{i=1}^{d} \frac{1}{x_i} \right) \exp \left(-\frac{1}{2} (\log x - \mu)^\top \Sigma^{-1} (\log x - \mu) \right), \quad x \in \mathbb{R}_+^d, \tag{34}$$

where log is applied element-wise. Namely, log-transformed data is assumed to follow the usual Gaussian graphical model. Note that this model is also an exponential family and thus the objective function of score matching reduces to a quadratic form (8). Here, we apply SMIC to determine which of the above two graphical models has better fit to RNAseq data.

Figure 5 plots SMIC of two graphical models with respect to the number of edges. For edge selection, we employed l_1 regularized score matching (Lin et al., 2016) for truncated Gaussian graphical models (32) and graphical LASSO[7] for log-Gaussian graphical models (34), respectively. Namely, we computed the whole regularization paths. After edge selection, we fitted the graphical models again by score matching without regularization to calculate SMIC. Note that such a procedure is also used for LASSO (Belloni and Chernozhukov, 2013). Figure 5 indicates that SMIC saturates around 400 edges for both models and is smaller for the log-Gaussian graphical model. Therefore, the log-Gaussian graphical model has better fit to RNAseq data in this case.

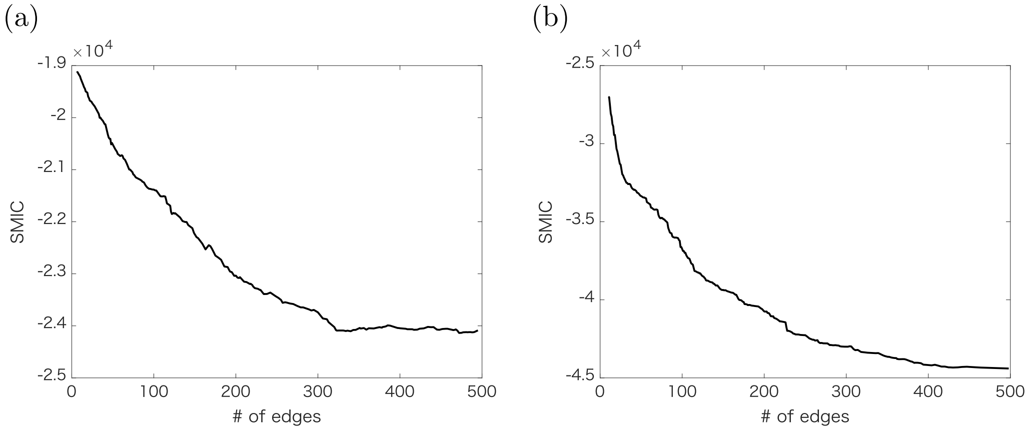

Figure 5: SMIC of (a) truncated Gaussian graphical models (32) and (b) log-Gaussian graphical models (34) for RNAseq data. Note that the scale of y-axis is different between (a) and (b).

8.3 Wind direction data

Finally, we apply NCIC to wind direction data. Since the wind direction is naturally identified with a vector on the unit circle (Mardia and Jupp, 2008), we represent it as a circular variable in radians. Figure 6 shows a 2-d histogram of wind direction at Tokyo on 00:00 (x_1) and 12:00 (x_2) for $N = 365$ days in 2018, which was obtained from the website of Japan Meteorological Agency. The data are discretized into 16 bins such as north-northeast.

To describe dependence between two circular variables, Singh et al. (2002) proposed the bivariate von Mises distribution defined by

$$p(x_1, x_2 \mid \theta) \propto \exp(\kappa_1 \cos(x_1 - \mu_1) + \kappa_2 \cos(x_2 - \mu_2) + \lambda_{12} \sin(x_1 - \mu_1) \sin(x_2 - \mu_2)),(35)$$

where $\theta = (\kappa_1, \kappa_2, \mu_1, \mu_2, \lambda_{12})$ with $\kappa_1 \geq 0$, $\kappa_2 \geq 0$, $0 \leq \mu_1 < 2\pi$ and $0 \leq \mu_2 < 2\pi$. Its normalization constant involves an infinite sum of Bessel functions, which is computationally

[7] We used R package "glasso" from `http://statweb.stanford.edu/~tibs/glasso/`.

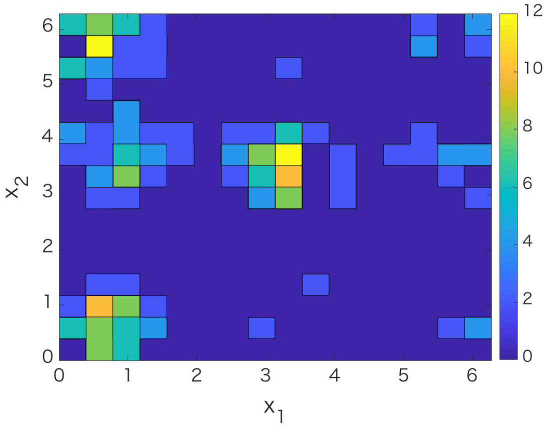

Figure 6: 2-d histogram of wind direction data.

intractable. The parameter λ_{12} quantifies the dependency between x_1 and x_2. In particular, x_1 and x_2 are independent if and only if $\lambda_{12} = 0$.

We fitted the bivariate von Mises distribution (35) to the wind direction data in Figure 6 by NCE with $M = 1000$ noise samples from the uniform distribution on $[0, 2\pi) \times [0, 2\pi)$. The parameter estimate was

$$(\hat{\kappa}_1, \hat{\kappa}_2, \hat{\mu}_1, \hat{\mu}_2, \hat{\lambda}_{12}) = (0.813, 0.440, 1.120, 4.644, -0.965)$$

and NCIC value was $\mathrm{NCIC}_2 = -1941$. We also fitted the bivariate von Mises distribution (35) with $\lambda_{12} = 0$. The parameter estimate was

$$(\hat{\kappa}_1, \hat{\kappa}_2, \hat{\mu}_1, \hat{\mu}_2) = (0.808, 0.430, 0.755, 4.234)$$

and NCIC value was $\mathrm{NCIC}_2 = -1919$. Thus, the former model has better fit than the latter, which implies that the wind direction at Tokyo on 00:00 and 12:00 are dependent.

9. Extension to non-normalized mixture models

In this section, we discuss extension of NCIC to non-normalized mixture models.

Consider a finite mixture of non-normalized models:

$$p(x \mid \theta, \pi) = \sum_{k=1}^{K} \pi_k \cdot p(x \mid \theta_k), \quad p(x \mid \theta_k) = \frac{1}{Z(\theta_k)} \widetilde{p}(x \mid \theta_k), \tag{36}$$

where $\pi_k > 0$, $\sum_{k=1}^{K} \pi_k = 1$ and the normalization constant $Z(\theta_k)$ of each component $p(x \mid \theta_k)$ is intractable. Existing methods for estimating non-normalized models are not applicable to (36) since it includes more than one intractable normalization constant[8]. We clarified this point as a footnote. Thus, Matsuda and Hyvärinen (2019) extended NCE to

[8] For example, a formal extension of score matching becomes intractable because the objective function now involves the normalization constants.

estimate (36). Specifically, (36) is reparametrized as

$$p(x \mid \theta, c) = \sum_{k=1}^{K} p(x \mid \theta_k, c_k), \quad \log p(x \mid \theta_k, c_k) = \log \widetilde{p}(x \mid \theta_k) + c_k, \tag{37}$$

where $c = (c_1, \dots, c_K)$ with $c_k = \log \pi_k - \log Z(\theta_k)$. Similarly to the original NCE, we consider c as an additional unknown parameter. Then, by generating M noise samples $y^{(1)}, \dots, y^{(M)}$ from a noise distribution $n(y)$, the parameter $\xi = (\theta, c)$ is estimated in the same way as the original NCE in (3) and (4), that is, we use the definition (37) in the original NCE objective function (4). This extended NCE has consistency under mild regularity conditions (Matsuda and Hyvärinen, 2019).

Now, we consider extension of NCIC to non-normalized mixture models. The problem setting is essentially the same with Section 5. Specifically, we have N i.i.d. samples $x^{(1)}, \dots, x^{(N)}$ from an unknown distribution $q(x)$ and estimate a non-normalized mixture model (37) by using the extended NCE. Assume that the distribution $p_*(x) = p(x \mid \xi^*)$ with $\xi^* = \arg\min_{\xi} d_{\mathrm{NCE}}(q, p_{\xi})$ has exactly K mixture components: $\pi_1^* > 0, \dots, \pi_K^* > 0$ and $\theta_i^* \neq \theta_j^*$ $(i \neq j)$. In this case, the model is regular around ξ^*. Therefore, Lemma 2 is valid and so NCIC_1 in (23) is approximately unbiased. Also, by replacing $j_m(\xi^*)$ with $h = \sum_{l=m-K+1}^{m} j_l(\xi^*)$ in the proof, Lemma 3 for well-specified cases is valid as well, where the value of m is changed from Section 5 to $m = \dim(\xi) = K(\dim(\theta_1) + 1)$. Therefore, we propose

$$\mathrm{NCIC}_2 = N \hat{d}_{\mathrm{NCE}}(\hat{\xi}_{\mathrm{NCE}}) + K \left\{ \dim(\theta_1) + 1 \right\} - \frac{1}{N+M} \left\{ \sum_{t=1}^{N} \hat{b}(x^{(t)}) + \sum_{t=1}^{M} \hat{b}(y^{(t)}) \right\}$$

as an approximately unbiased estimator of $N \mathrm{E}_{x,y} \{ d_{\mathrm{NCE}}(q, \hat{p}) \}$, where $\hat{b}(z)$ is defined as (24). Thus, we can select the number of components K of non-normalized mixture models (37) by minimizing NCIC.

Figure 7 shows a result on the non-normalized version of the Gaussian mixture distribution:

$$p(x \mid \theta, c) = \sum_{k=1}^{K} \exp(\theta_{k1} x^2 + \theta_{k2} x + c_k). \tag{38}$$

Here, we generated $N = 10^3$ samples from the two-component Gaussian mixture distribution $0.5 \cdot \mathrm{N}(0,1) + 0.5 \cdot \mathrm{N}(3,1)$ and applied the extended NCE to estimate (38). The noise distribution was set to the Gaussian distribution with the same mean and variance as data and the noise sample size was set to $M = 10^4$. For AIC, we computed the maximum likelihood estimator with the MATLAB function *fitgmdist*. Both NCIC_2 and AIC take minimum at the true value $K = 2$.

In the above, we assumed that the model is regular around ξ^*. It is not trivial to eliminate this condition due to the singularity in the parameter space of finite mixture models (Mclachlan and Peel, 2004). It is an interesting future work to develop a rigorous theory of model selection for non-normalized mixture models accounting for singularity (Watanabe, 2021).

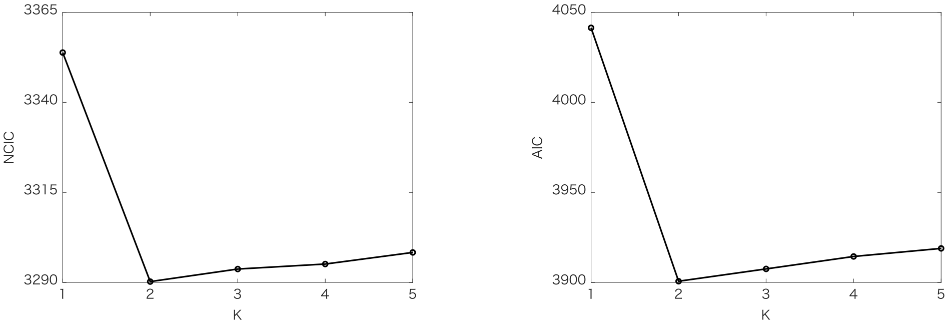

Figure 7: NCIC_2 (left) and AIC (right) for Gaussian mixture models. The true value is $K = 2$.

10. Conclusion

In this study, we developed information criteria for non-normalized models estimated by noise contrastive estimation (NCE) or score matching. The proposed criteria are approximately unbiased estimators of discrepancy measures for non-normalized models. They provide a principled method of model selection for general non-normalized models. We believe that this study increases the practicality of non-normalized models.

Regarding future work, an interesting direction would be to apply NCIC to data-driven selection of neural network architectures (Murata et al., 1994), which is a constant problem in deep learning. In a sense, the experiment of overcomplete independent component analysis on natural image data in Section 8.1 is viewed as selecting the number of units. In a similar way, NCIC may be applicable to select the number of layers. Note that Gutmann and Hyvärinen (2012) applied NCE to train neural networks on natural image data. It would be also interesting if we can select from different architectures such as ResNet and CNN. Furthermore, since regularization is essential to avoid overfitting in training high-dimensional models including neural networks, it is an important problem to extend NCIC and SMIC to regularized cases such as LASSO (Ninomiya and Kawano, 2016). On the other hand, since recent studies (Fujikoshi et al., 2014; Yanagihara et al., 2015; Bai et al., 2018) have found that AIC attains consistency in high-dimensional settings (in contrary to low-dimensional settings), it is an interesting future work to investigate the consistency of NCIC and SMIC in high-dimensional settings.

Acknowledgments

We thank Yoshiyuki Ninomiya and Ricardo Monti for helpful comments. T.M. was supported by JSPS KAKENHI Grant Number 19K20220 and JST Moonshot Grant Number JPMJMS2024. A.H. was supported by a Fellowship from CIFAR.

References

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

Bai, Z., Choi, K. P. & Fujikoshi, Y. (2018). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *The Annals of Statistics*, **46**, 1050–1076.

Barp, A., Briol, F. X., Duncan, A., Girolami, M. & Mackey, L. (2019). Minimum Stein discrepancy estimators. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019).*

Belloni, A. & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, **19**, 521–547.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, **36**, 192–236.

Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference.* New York: Springer.

Caimo, A. & Friel, N, J. (2011). Bayesian inference for exponential random graph models. *Social Networks*, **33**, 41–55.

Chikuse, Y. (2003). *Statistics on Special Manifolds.* New York: Springer.

Claeskens, G. & Hjort, N. L. (2008). *Model Selection and Model Averaging.* Cambridge: Cambridge University Press.

Dawid, A. P. & Musio, M. (2015). Bayesian model selection based on proper scoring rules. *Bayesian Analysis*, **10**, 479–499.

Drton, M. & Perlman, M. D. (2004). Model selection for Gaussian concentration graphs. *Biometrika*, **91**, 591–602.

Forbes, P. G. M. & Lauritzen, S. (2015). Linear estimating equations for exponential families with application to Gaussian linear concentration models. *Linear Algebra and its Applications*, **473**, 261–283.

Fujikoshi, Y., Sakurai, T. & Yanagihara, H. (2014). Consistency of high-dimensional AIC-type and Cp-type criteria in multivariate linear regression. *Journal of Multivariate Analysis*, **123**, 184–200.

Gelman, A., Hwang, J. & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, **24**, 997–1016.

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society B*, **56**, 261–274.

Grant, M. & Boys, S. (2018). CVX: Matlab software for disciplined convex programming. version 2.1, December 2018. `http://cvxr.com/cvx`

Gutmann, M. U. & Hirayama, J. (2011). Bregman divergence as general framework to estimate unnormalized statistical models. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011).*

Gutmann, M. U. & Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the 13th International Workshop on Artificial Intelligence and Statistics (AISTATS 2010).*

Gutmann, M. U. & Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, **13**, 307–361.

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

Hyvärinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis.* New York: Wiley.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6**, 695–709.

Hyvärinen, A. (2007). Some extensions of score matching. *Computational Statistics & Data Analysis*, **51**, 2499–2512.

Hyvärinen, A., Hurri, J. & Hoyer, P. O. (2009). *Natural Image Statistics: A probabilistic approach to early computational vision.* New York: Springer.

Ji, C. & Seymour, L. (1996). A consistent model selection procedure for Markov random fields based on penalized pseudolikelihood. *Annals of Applied Probability*, **6**, 423–443.

Kitagawa, G. (1997). Information criteria for the predictive evaluation of bayesian models. *Communications in Statistics - Theory and Methods*, **26**, 2223–2246.

Konishi, S. & Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875–890.

Konishi, S. & Kitagawa, G. (2008). *Information Criteria and Statistical Modeling.* New York: Springer.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford: Oxford University Press.

Li, S. Z. (2001). *Markov Random Field Modeling in Image Analysis.* New York: Springer.

Lin, L., Drton, M. & Shojaie, A. (2016). Estimation of high-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics*, **10**, 806–854.

Liu, S., Kanamori, T., Jitkrittum, W. & Chen, Y. (2019). Fisher efficient inference of intractable models. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019).*

Lyu, S. (2009). Interpretation and Generalization of Score Matching. In *Proceedings of the 25th International Conference on Uncertainty in Artificial Intelligence (UAI 2009).*

Mardia, K. V. & Jupp, P. E. (2008). *Directional Statistics.* New York: Wiley.

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

Mattheou, K., Leeb, S. & Karagrigoriou, A. (2001). A model selection criterion based on the BHHJ measure of divergence. *Journal of Statistical Planning and Inference*, **139**, 228–235.

Mclachlan, G. & Peel, D. (2000). *Finite Mixture Models.* New York: Wiley.

Murata, N., Yoshizawa, S. & Amari, S. (1994). Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, **5**, 865–872.

Ninomiya, Y. & Kawano, S. (2016). AIC for the Lasso in generalized linear models. *Electronic Journal of Statistics*, **10**, 2537–2560.

Olshausen, B. A. & Field, D. J. (2001). Sparse coding with an overcomplete basis set: A strategy employed by V1?. *Vision Research*, **37**, 3311–3325.

Pan, W. (2001). Akaike's Information Criterion in Generalized Estimating Equations. *Biometrics*, **57**, 120–125.

Parry, M., Dawid, A. P. & Lauritzen, S. (2012). Proper local scoring rules. *The Annals of Statistics*, **40**, 561–592.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, **85**, 619–630.

Rasmussen, C. E. (2006). Conjugate gradient algorithm. Matlab code version 2006-09-08. `http://learning.eng.cam.ac.uk/carl/code/minimize/minimize.m`

Ravikumar, P., Wainwright, M. J. & Lafferty, J. D. (2010). High-dimensional Ising model selection using l_1-regularized logistic regression. *The Annals of Statistics*, **38**, 1287–1319.

Riou-Durand, L. & Chopin, N. (2018). Noise contrastive estimation: Asymptotic properties, formal comparison with MC-MLE. *Electronic Journal of Statistics*, **12**, 3473–3518.

Shao, S., Jacob, P. E., Ding, J. & Tarokh, V. (2019). Bayesian model comparison with the Hyvarinen score: computation and consistency. *Journal of the American Statistical Association*, **114**, 1826–1837.

Singh, H., Hnizdo, V. & Demchuk, E. (2002). Probabilistic model for two dependent circular variables. *Biometrika*, **89**, 719–723.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, **64**, 583–639.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society B*, **39**, 44–47.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Sciences*, **153**, 12–18 (in Japanese).

Teh, Y., Welling, M., Osindero, S. & Hinton, G. E. (2004). Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, **4**, 1235–1260.

Uehara, M., Matsuda, T. & Komaki, F. (2018). Analysis of noise contrastive estimation from the perspective of asymptotic variance. arXiv:1808.07983.

Uehara, M., Kanamori, T., Takenouchi, T. & Matsuda, T. (2020a). A unified statistically efficient estimation framework for unnormalized models. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020).*

Uehara, M., Matsuda, T. & Kim, J. K. (2020b). Imputation estimators for unnormalized models with missing data. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020).*

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

Varin, C. & Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, **92**, 519–528.

Watanabe, S. (2021). WAIC and WBIC for mixture models. *Behaviormetrika*, **48**, 5–21.

Watanabe, S. & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**, 3571–3594.

Wooldridge, J. M. (2001). Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometric Theory*, **17**, 451–470.

Yanagihara, H., Wakaki, H. & Fujikoshi, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electronic Journal of Statistics*, **9**, 869–897.

Yu, S., Drton, M. & Shojaie, A. (2018). Graphical models for non-negative data using generalized score matching. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS 2018).*

Yu, S., Drton, M. & Shojaie, A. (2019). Generalized score matching for non-negative data. *Journal of Machine Learning Research*, **20**, 1–70.