entropy

MDPI

Article

Efficiency Bound of Local Z-Estimators on Discrete Sample Spaces

Takafumi Kanamori

Department of Computer Science and Mathematical Informatics, Nagoya University, Furocho, Chikusaku, Nagoya 464-8603, Japan; kanamori@is.nagoya-u.ac.jp; Tel.: +81-52-789-4598

Academic Editor: Kevin H. Knuth
Received: 14 June 2016; Accepted: 20 July 2016; Published: 23 July 2016

Abstract: Many statistical models over a discrete sample space often face the computational difficulty of the normalization constant. Because of that, the maximum likelihood estimator does not work. In order to circumvent the computation difficulty, alternative estimators such as pseudo-likelihood and composite likelihood that require only a local computation over the sample space have been proposed. In this paper, we present a theoretical analysis of such localized estimators. The asymptotic variance of localized estimators depends on the neighborhood system on the sample space. We investigate the relation between the neighborhood system and estimation accuracy of localized estimators. Moreover, we derive the efficiency bound. The theoretical results are applied to investigate the statistical properties of existing estimators and some extended ones.

Keywords: Z-estimator; stochastic localization; efficiency; pseudo-likelihood; composite likelihood

1. Introduction

For many statistical models on a discrete sample space, the computation of the normalization constant is often intractable. Because of that, the maximum likelihood estimator (MLE) is not of practical use to estimate probability distributions, although the MLE has some nice theoretical properties such as the statistical consistency and efficiency under some regularity conditions [1].

In order to circumvent the computation of the normalization constant, alternative estimators that require only a local computation over the sample space have been proposed. In this paper, estimators on the basis of such a concept are called *localized estimators*. Examples of localized estimators include pseudo-likelihood [2], composite likelihood [3,4], ratio matching [5,6], proper local scoring rules [7,8], and many others. These estimators are used for discrete statistical models such as conditional random fields [9], Boltzmann machines [10], restricted Boltzmann machines [11], discrete exponential family harmoniums [12], and Ising models [13].

In this paper, we present a theoretical analysis of localized estimators. We use the standard tools in the statistical asymptotic theory. In our analysis, a class of localized estimators including pseudo-likelihood and composite likelihood is treated as M-estimator or Z-estimator which is an extension of the MLE [1]. The localized estimators require local computation around a neighborhood of observed points. Hence, the asymptotic variance of the localized estimator depends on the size of the neighborhood. We investigate the relation between the estimation accuracy and the neighborhood system. A similar result is given by [14], in which asymptotic variances between specific composite likelihoods are compared. In our approach, we consider a stochastic variant of localized estimators, and derive a general result that the larger neighborhood leads to more efficient estimator under a simple condition. The pseudo-likelihood and composite likelihood are obtained as the expectation of a stochastic localized estimator. We derive the exact efficiency bound for the expected localized estimator. As far as we know, the derivation of the efficiency bound is a new result

for a class of localized estimators, though upper and lower bounds have been proposed [14] for some localized estimators.

The rest of the paper is organized as follows. In Section 2, we introduce basic concepts such as pseudo-likelihood, composite likelihood, and Z-estimators. Section 3 is devoted to define stochastic local Z-estimator associated with a neighborhood system over the discrete sample space. In Section 4, we study the relation between the neighborhood system and asymptotic efficiency of the stochastic local Z-estimator. In Section 5, we define local Z-estimator as the expectation of the stochastic local Z-estimator, and present its efficiency bound. The theoretical results are applied to study asymptotic properties of existing estimators and some extended ones. Finally, Section 6 concludes the paper with discussions.

2. Preliminaries

M-estimators and Z-estimators were proposed as an extension of the MLE. In practice, M-estimators and Z-estimators are often computationally demanding due to the normalization constant in statistical models. To circumvent computational difficulty, localized estimators have been proposed. We introduce some existing localized estimators especially on discrete sample spaces. In later sections, we consider statistical properties of a localized variant of Z-estimators.

Let us summarize the notations to be used throughout the paper. Let $\mathbb{R}$ be the set of all real numbers. The discrete sample space is denoted as $\mathcal{X}$. The statistical model $p_\theta(x)$ for $x \in \mathcal{X}$ with the parameter $\theta \in \Theta \subset \mathbb{R}^d$ is also expressed as $p(x;\theta)$. The vector $\boldsymbol{a}$ usually denotes the column vector, and $\cdot^T$ denotes the transposition of vector or matrix. For a linear space T and an integer d, $(T)^d$ denotes the d-fold product space of T, and the element $\boldsymbol{c} \in (T)^d$ is expressed as $\boldsymbol{c} = (c_1, \ldots, c_d)$. The product space of two subspaces T_1 and T_2 that are orthogonal to each other is denoted as $T_1 \oplus T_2$. For the function $f(\theta)$ of the parameter $\theta \in \mathbb{R}^d$, ∇f denotes the gradient vector $(\frac{\partial f}{\partial \theta_1}, \ldots, \frac{\partial f}{\partial \theta_d})^T$. The indicator function is denoted as $\mathbf{1}[A]$ that takes 1 if A is true and 0 otherwise.

2.1. M- and Z-Estimators

Suppose that samples $x_1, \ldots, x_m$ are i.i.d. distributed from the probability $p(x)$ over the discrete sample space $\mathcal{X}$. A statistical model $p_\theta(x) = p(x;\theta)$ with the parameter $\theta \in \Theta \subset \mathbb{R}^d$ is assumed to estimate $p(x)$. In this paper, our concern is the statistical efficiency of estimators. Hence, we suppose that the statistical model includes $p(x)$.

The MLE is commonly used to estimate $p(x)$. It uses the negative log-likelihood of the model, $-\log p_\theta(x)$, as the loss function and the estimator is given by the minimum solution of its empirical mean, $-\frac{1}{m}\sum_{i=1}^m \log p_\theta(x_i)$.

Generally, the estimator obtained by the minimum solution of a loss function is referred to as the M-estimator. The MLE is an example of M-estimators. When the loss function is differentiable, the gradient of the loss function vanishes at the estimated parameter. Instead of minimizing loss function, a solution of the system of equations also provides an estimator of the parameter. Such an estimator is called the Z-estimator [1]. In the MLE, the system of equations is given as

$$\frac{1}{m}\sum_{i=1}^m \nabla \log p_\theta(x_i) = \mathbf{0},$$

where $\mathbf{0} \in \mathbb{R}^d$ is the null-vector. The gradient $\nabla \log p_\theta(x)$ is known as the *score function* of the model $p_\theta(x)$. In this paper, the score function is denoted as

$$\boldsymbol{u}_\theta(x) = \nabla \log p_\theta(x).$$

In general, the Z-estimator is defined as the solution of the system of equations

$$\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{f}_\theta(x_i) = \mathbf{0},$$

where the $\mathbb{R}^d$-valued function $\boldsymbol{f}_\theta(x) = \boldsymbol{f}(x;\theta)$ is referred to as the *identification function* [15,16]. In the M-estimator, the identification function is given as the gradient of the loss function. In general, however, the identification function itself is not necessarily expressed as the gradient of a loss function, if it is not integrable. The identification function $\boldsymbol{f}_\theta(x)$ is also called Z-estimator with some abuse of terminology.

2.2. Localized Estimators

Below, let us introduce some localized estimators. The statistical model defined on the discrete set $\mathcal{X}$ is denoted by

$$p_\theta(x) = \frac{\widetilde{p}_\theta(x)}{Z_\theta} \tag{1}$$

for $x \in \mathcal{X}$, where Z_θ is the normalization constant at the parameter θ, i.e.,

$$Z_\theta = \sum_{x\in\mathcal{X}} \widetilde{p}_\theta(x).$$

Throughout the paper, we assume $p_\theta(x) > 0$ for all $x \in \mathcal{X}$ and all $\theta \in \Theta \subset \mathbb{R}^d$.

Example 1 (Pseudo-likelihood)**.** Suppose that $\mathcal{X}$ is expressed as the product space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, where $\mathcal{X}_1, \ldots, \mathcal{X}_n$ are finite sets such as $\{0,1\}$. For $x = (x_1, \ldots, x_n) \in \mathcal{X}$, let x_{-k} be the $n-1$ dimensional vector defined by dropping the k-th element of x. The loss function of the pseudo-likelihood, S_{PS}, is defined as the negative log-likelihood of the conditional probability $p_\theta(x_k|x_{-k})$ defined from $p_\theta(x)$, i.e.,

$$S_{\mathrm{PS}}(x, p_\theta) = -\sum_{k=1}^{n} \log p_\theta(x_k|x_{-k}) = -\sum_{k=1}^{n} \Big\{ \log \widetilde{p}_\theta(x) - \log \Big(\sum_{x_k\in\mathcal{X}_k} \widetilde{p}_\theta(x)\Big)\Big\}. \tag{2}$$

The pseudo-likelihood does not require the normalization constant, and it satisfies the statistical consistency of the parameter estimation [2,17]. The identification function of the corresponding Z-estimator is obtained by the gradient vector of the loss Function (2).

Example 2 (Composite likelihood)**.** The composite likelihood was proposed as an extension of the pseudo-likelihood [3]. Suppose that $\mathcal{X}$ is expressed as the product space as in Example 1. For the index subset $A \subset \{1, \ldots, n\}$, let $x_A = (x_i)_{i\in A}$ be the subvector of $x \in \mathcal{X}$. For each $\ell = 1, \ldots, M$, Suppose that A_ℓ and B_ℓ are a pair of disjoint subsets in $\{1, \ldots, n\}$, and let C_ℓ be the complement of the union $A_\ell \cup B_\ell$, i.e., $C_\ell = (A_\ell \cup B_\ell)^c$. Given positive constants $\gamma_1, \ldots, \gamma_M$, the loss function of the composite likelihood, S_{CL}, is defined as

$$S_{\mathrm{CL}}(x, p_\theta) = -\sum_{\ell=1}^{M} \gamma_\ell \log p_\theta(x_{A_\ell}|x_{B_\ell}) = -\sum_{\ell=1}^{M} \gamma_k \log \Big\{ \sum_{x_{C_\ell}} \widetilde{p}_\theta(x) - \sum_{x_{B_\ell^c}} \widetilde{p}_\theta(x)\Big\}.$$

The composite likelihood using the subsets $A_\ell = \{\ell\}$, $B_\ell = A_\ell^c$ and positive constant $\gamma_\ell = 1$ for $\ell = 1, \ldots, n$ yields the pseudo-likelihood. As well as the pseudo-likelihood, the composite likelihood has the statistical consistency under some regularity condition [4].

Originally, the pseudo and composite likelihoods were proposed to deal with spatial data [2,3]. As a generalization of these estimators, a localized variant of scoring rules works efficiently to the statistical analysis of discrete spatial data [18].

3. A Stochastic Variant of Z-Estimators

In this section, we define a stochastic variant of Z-estimators. For the discrete sample space $\mathcal{X}$, suppose that the neighborhood system N is defined as a subset of the power set $2^{\mathcal{X}}$, i.e., N is a family of subsets in $\mathcal{X}$. Let us define the neighborhood system at $x \in \mathcal{X}$ by $N_x = \{e \in N | x \in e\}$. We assume that N_x is not empty for any x. In some class of localized estimators, the neighborhood system is expressed using an undirected graph on $\mathcal{X}$ [7]. In our setup, the neighborhood system is not necessarily expressed by an undirected graph, and we allow the neighborhood system to possess multiple neighbors at each point x.

Let us define the stochastic Z-estimator. A conditional probability of the set $e \in N$ given $x \in \mathcal{X}$ is denoted as $q(e|x)$. We assume that $q(e|x) = 0$ if $e \notin N_x$ throughout the paper. Given a sample x, we randomly generate a neighborhood e from the conditional probability $q(e|x)$. Using i.i.d. copies of (x, e), we estimate $p(x)$. Here, the statistical model $p_\theta(x)$ of the form (1) is used. We use the Z-estimator $\boldsymbol{f}_\theta(x, e) = \boldsymbol{f}(x, e; \theta) \in \mathbb{R}^d$ to estimate the parameter $\theta \in \Theta \subset \mathbb{R}^d$. The element of $\boldsymbol{f}_\theta(x, e)$ is denoted as $f_{\theta,k}(x, e)$ or $f_k(x, e; \theta)$ for $k = 1, \ldots, d$. The expectation under the probability $p_\theta(x)q(e|x)$ is written as $\mathbb{E}_{\theta,q}[\cdot]$. Suppose that the equality

$$\mathbb{E}_{\theta,q}[\boldsymbol{f}_\theta] = \boldsymbol{0} \tag{3}$$

holds for all $\theta \in \Theta$. In addition, we assume that the vectors $\mathbb{E}_{\theta,q}[\nabla f_{\theta,k}]$, $k = 1, \ldots, d$ are linearly independent, meaning that $\boldsymbol{f}_\theta$ depends substantially on the parameter θ [19]. The solution of the system of equations

$$\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{f}_\theta(x_i, e_i) = \boldsymbol{0} \tag{4}$$

produces a statistically consistent estimator under some regularity condition [1]. In the parameter estimation of the model $p_\theta(x)$, the *stochastic Z-estimator* is defined as the solution of (4) using the identification function satisfying (3). As shown in Section 5, stochastic Z-estimators are useful to investigate statistical properties of the standard pseudo-likelihood and composite likelihood in Examples 1 and 2.

The computational tractability of the stochastic Z-estimator is not necessarily guaranteed. The MLE using the score function $\boldsymbol{f}_\theta(x, e) = \boldsymbol{u}_\theta(x)$ is regarded as a stochastic Z-estimator for any $q(e|x)$ and it may not be computationally tractable because of the normalizing constant. As a class of computationally efficient estimators, let us define the *stochastic local Z-estimator* as the stochastic Z-estimator using $\boldsymbol{f}_\theta(x, e)$ satisfying

$$\mathbb{E}_{\theta,q}[\boldsymbol{f}_\theta | e] = \boldsymbol{0} \tag{5}$$

for any neighborhood $e \in N$, where $\mathbb{E}_{\theta,q}[\cdot|e]$ is the conditional expectation given e. The conditional probability $p(x|e)$ of $p_\theta(x)q(e|x)$ can take a positive value only when $x \in e$. Hence, $\boldsymbol{f}_\theta(x, e)$ depends only on the neighborhood of x and its computation will be tractable.

Example 3 (Stochastic pseudo-likelihood)**.** Let us define the stochastic variant of the pseudo-likelihood estimator. On the sample space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, the neighborhood system N_x at $x = (x_1, \ldots, x_n) \in \mathcal{X}$ is defined as $N_x = \{e_{x,k} \,|\, k = 1, \ldots, n\}$, where $e_{x,k} \subset \mathcal{X}$ is given as

$$e_{x,k} = \{(x_1, \ldots, x_{k-1}, z_k, x_{k+1}, \ldots, x_n) \,|\, z_k \in \mathcal{X}_k\}.$$

In order to estimate the parameters in complex models such as conditional random fields and Boltzmann machines, the union, $\cup_{e\in N_x} e$, is often used as the neighborhood at x [2,9,17]. Let the conditional probability $q(e|x)$ on N_x as $q(e_{x,k}|x) = q_k$, $k = 1,\dots,n$, where $q_1,\dots,q_n$ are positive numbers satisfying $\sum_{k=1}^n q_k = 1$. The identification function of the stochastic pseudo-likelihood is defined by

$$\boldsymbol{f}_\theta(x,e) = \nabla \log \frac{p_\theta(x)}{\sum_{z\in e} p_\theta(z)}$$

for $e \in N_x$. Then, $\boldsymbol{f}_\theta(x,e_{x,k})$ is equal to $\nabla \log p_\theta(x_k|x_{-k})$. The conditional probability $p(x|e)$ derived from $p_\theta(x)q(e|x)$ is given as

$$p(x|e_{x,k}) = \frac{p_\theta(x)q(e_{x,k}|x)}{\sum_{z\in e_{x,k}} p_\theta(z)q(e_{x,k}|z)} = \frac{p_\theta(x)q_k}{\sum_{z\in e_{x,k}} p_\theta(z)q_k} = p_\theta(x_k|x_{-k}),$$

where we used the equality $e_{x,k} = e_{z,k}$ for $z \in e_{x,k}$. Hence, the equality (5) holds for any $(q_k)_{k=1,\dots,n}$. When q_k depends on x, $p(x|e_{x,k})$ is different from $p_\theta(x_k|x_{-k})$ in general.

Example 4 (Stochastic variant of composite likelihood)**.** Let us introduce a stochastic variant of composite likelihood on the sample space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. Below, notations in Example 2 are used. Let us define $e'_{x,\ell}$, $\ell = 1,\dots,M$ by the subset $e'_{x,\ell} = \{y \in \mathcal{X} \,|\, y_{B_\ell} = x_{B_\ell}\}$, and the neighborhood system N'_x by $N'_x = \{e'_{x,\ell} \,|\, \ell = 1,\dots,M\}$. We assume that the map from ℓ to B_ℓ is one to one. In other words, the disjoint subsets A_ℓ, B_ℓ, C_ℓ can be specified from the neighborhood $e'_{x,\ell}$. Suppose that the conditional probability $q'(e'|x)$ on N'_x is defined as $q'(e'_{x,\ell}|x) = q'_\ell$ for $\ell = 1,\dots,M$, where $q'_1,\dots,q'_M$ are positive numbers satisfying $\sum_{\ell=1}^M q'_\ell = 1$. As well as Example 3, we see that the conditional probability $q'(x|e'_{x,\ell})$ defined from $p_\theta(x)q'(e'|x)$ is given as $p_\theta(x_{A_\ell}, x_{C_\ell}|x_{B_\ell})$. Let us consider the identification function,

$$\boldsymbol{f}_\theta(x,e'_{x,\ell}) = \nabla \log \frac{\sum_{x_{C_\ell}} p_\theta(x)}{\sum_{x_{B_\ell^c}} p_\theta(z)}, \tag{6}$$

which is nothing but $\nabla \log p_\theta(x_{A_\ell}|x_{B_\ell})$. Then, (5) holds under the joint probability $p_\theta(x)q'(e'|x)$. Indeed, we have

$$\mathbb{E}_{\theta,q'}[\boldsymbol{f}_\theta|e'_{x,\ell}] = \sum_{x_{A_\ell}, x_{C_\ell}} p_\theta(x_{A_\ell}, x_{C_\ell}|x_{B_\ell}) \nabla \log p_\theta(x_{A_\ell}|x_{B_\ell}) = \boldsymbol{0}$$

for any $(q'_\ell)_{\ell=1,\dots,M}$. In this paper, the Z-estimator using (6) is called the reduced stochastic composite likelihood (reduced-SCL). The stochastic composite likelihood proposed in [20] is a randomized extension of the above $\boldsymbol{f}_\theta(x,e')$. Let $z = (z_1,\dots,z_M)$ be a binary random vector taking an element of $\{0,1\}^M$, and $\alpha_1,\dots,\alpha_M$ be positive constants. The SCL is defined as the Z-estimator obtained by

$$\boldsymbol{f}(x,z;\theta) = \sum_{\ell=1}^M \alpha_\ell z_\ell \nabla \log p(x_{A_\ell}|x_{B_\ell};\theta).$$

The statistical consistency and the normality of the SCL is shown in [20].

4. Neighborhood Systems and Asymptotic Variances

We consider the relation between neighborhood systems and statistical properties of stochastic local Z-estimators.

4.1. Tangent Spaces of Statistical Models

At the beginning, let us introduce some geometric concepts to investigate statistical properties of localized estimators. These concepts are borrowed from information geometry [21]. For the neighborhood system N with the conditional probability $q(e|x)$, let us define the linear space $T_{\theta,q}$ as

$$T_{\theta,q} = \{a : \mathcal{X} \times N \to \mathbb{R} \,|\, \mathbb{E}_{\theta,q}[a] = 0,\ a(x,e) = 0 \text{ if } q(e|x) = 0\}.$$

The inner product for $a_1, a_2 \in T_{\theta,q}$ is defined as $\mathbb{E}_{\theta,q}[a_1 a_2]$. A geometric meaning of $T_{\theta,q}$ is the tangent space of the statistical model $\{p_\theta(x)q(e|x)|\theta \in \Theta\}$. For any $a \in T_{\theta,q}$ and sufficiently small $\varepsilon > 0$, the perturbation of $p_\theta(x)q(e|x)$ to the direction $a(x,e)$ leads to the probability function $p_\theta(x)q(e|x)(1 + \varepsilon a(x,e))$. Each element of the score function $u_{\theta,j}(x), j = 1, \ldots, d$ is a member of $T_{\theta,q}$ by regarding as $u_{\theta,j}(x) \cdot \mathbf{1}[q(e|x) \neq 0]$.

Let us consider the stochastic Z-estimator derived from $\boldsymbol{f}_\theta = (f_{\theta,1}, \ldots, f_{\theta,k})$ satisfying $\boldsymbol{f}_\theta \in (T_{\theta,q})^d$ for any θ. It leads to a Fisher consistent estimator. Stochastic local Z-estimators use an identification function in the linear subspace

$$T^L_{\theta,q} = \{f \in T_{\theta,q} \,|\, \mathbb{E}_{\theta,q}[f|e] = 0,\ \forall e \in N\}. \tag{7}$$

The orthogonal complement of $T^L_{\theta,q}$ in $T_{\theta,q}$ is denoted as $T^E_{\theta,q}$, which is given as

$$T^E_{\theta,q} = \{f \in T_{\theta,q} \,|\, f(x,e) \text{ does not depend on } x\}.$$

Indeed, the orthogonality of $T^L_{\theta,q}$ and $T^E_{\theta,q}$ is confirmed by

$$\mathbb{E}_{\theta,q}[a(x,e)b(e)] = \mathbb{E}_{\theta,q}[b(e)\mathbb{E}[a(x,e)|e]] = 0$$

for any $a \in T^L_{\theta,q}$ and any $b \in T^E_{\theta,q}$. In addition, any $f \in T_{\theta,q}$ can be decomposed into

$$f = (f - \mathbb{E}_{\theta,q}[f|e]) + \mathbb{E}[f|e]$$

such that $f - \mathbb{E}_{\theta,q}[f|e] \in T^L_{\theta,q}$ and $\mathbb{E}_{\theta,q}[f|e] \in T^E_{\theta,q}$.

The *efficient score* $\boldsymbol{u}^I_\theta = (u^I_{\theta,k})_{k=1,\ldots,d}$ is defined as the projection of each element of the score $\boldsymbol{u}_\theta$ onto $T^L_{\theta,q}$, i.e.,

$$\boldsymbol{u}^I_\theta(x,e) = \boldsymbol{u}_\theta(x) - \mathbb{E}_{\theta,q}[\boldsymbol{u}_\theta|e] = \nabla \log \widetilde{p}(x;\theta) - \mathbb{E}_{\theta,q}[\nabla \log \widetilde{p}(x;\theta)|e].$$

The efficient score is computationally tractable when the size of the neighborhood e is not exponential order but linear or low-degree polynomial order of n, where n is the dimension of x. The trade-off between the computational and statistical efficiency is presented in Theorems 1 and 2 in Section 4.2.

Another expression of the efficient score is

$$\boldsymbol{u}^I_\theta(x,e) = \nabla \log p(x|e;\theta,q), \tag{8}$$

where the conditional probability $p(x|e;\theta,q)$ is defined from $p_\theta(x)q(e|x)$. The above equality is obtained by

$$\begin{aligned}\boldsymbol{u}^I_\theta(x,e) &= \nabla \log p_\theta(x) - \sum_{x\in e} \frac{p_\theta(x)q(e|x)}{\sum_{x'} p_\theta(x')q(e|x')} \nabla \log p_\theta(x) \\ &= \nabla \log(p_\theta(x)q(e|x)) - \frac{\sum_{x\in e} \nabla p_\theta(x)q(e|x)}{\sum_{x'\in e} p_\theta(x')q(e|x')} \\ &= \nabla \log \frac{p_\theta(x)q(e|x)}{\sum_{x'} p_\theta(x')q(e|x')} = \nabla \log p(x|e;\theta,q).\end{aligned}$$

We define $T^I_{\theta,q}$ as the subspace of $T^L_{\theta,q}$ spanned by $\{u^I_{\theta,k} \,|\, k = 1,\dots,d\}$, and $T^A_{\theta,q}$ be the orthogonal complement of $T^I_{\theta,q}$ in $T^L_{\theta,q}$. As a result, we obtain

$$T_{\theta,q} = T^L_{\theta,q} \oplus T^E_{\theta,q}, \quad T^L_{\theta,q} = T^I_{\theta,q} \oplus T^A_{\theta,q}.$$

We describe statistical properties of stochastic local Z-estimators using the above tangent spaces.

4.2. Asymptotic Variance of Stochastic Local Z-Estimators

Under a fixed conditional probability $q(e|x)$, we derive the asymptotically efficient stochastic local Z-estimator in the same way to semi-parametric estimation [19,22]. In addition, we consider the monotonicity of the efficiency w.r.t. the size of the neighborhood. Given i.i.d. samples (x_i, e_i), $i = 1,\dots,m$, generated from $p(x)q(e|x)$, the estimator $\widehat{\theta}$ of the parameter in the model (1) is obtained by solving the system of Equation (4), where $\boldsymbol{f}_\theta \in (T_{\theta,q})^d$ for any $\theta \in \Theta$. Suppose that the true probability function $p(x)$ is realized by $p_\theta(x)$ of the model (1). As shown in [1], the standard asymptotic theory yields that the asymptotic variance of the above Z-estimator is given as

$$\lim_{n\to\infty} m \cdot \mathbb{V}[\widehat{\theta}] = \mathbb{E}_{\theta,q}[\boldsymbol{f}_\theta \boldsymbol{u}^T_\theta]^{-1} \mathbb{E}_{\theta,q}[\boldsymbol{f}_\theta \boldsymbol{f}^T_\theta] \mathbb{E}_{\theta,q}[\boldsymbol{u}_\theta \boldsymbol{f}^T_\theta]^{-1}. \tag{9}$$

The derivation of the asymptotic variance is presented in Appendix for completeness of the presentation.

We consider the asymptotic variance of the stochastic local Z-estimators. A simple expression of the asymptotic variance is obtained using the efficient score $\boldsymbol{u}^I_\theta$. Without loss of generality, the identification function of the stochastic local Z-estimator, $\boldsymbol{f}_\theta \in (T^L_{\theta,q})^d$, is expressed as

$$\boldsymbol{f}_\theta(x,e) = \boldsymbol{u}^I_\theta(x,e) + \boldsymbol{a}_\theta(x,e),$$

where $\boldsymbol{a}_\theta \in (T^A_{\theta,q})^d$. The reason is briefly shown below. Suppose that $\boldsymbol{f}_\theta(x,e)$ is decomposed into $\boldsymbol{f}_\theta(x,e) = B_\theta \boldsymbol{u}^I_\theta(x,e) + \boldsymbol{a}_\theta(x,e)$, where B_θ is a d by d matrix that does not depend on x and e. The condition that the matrix $\mathbb{E}_{\theta,q}[\nabla \boldsymbol{f}_\theta]$ is invertible assures that B_θ is invertible, since

$$\begin{aligned}\mathbb{E}_{\theta,q}\left[\frac{\partial}{\partial\theta_i} f_{\theta,k}\right] &= \sum_j \frac{\partial B_{\theta,kj}}{\partial\theta_i} \mathbb{E}_{\theta,q}\left[u^I_{\theta,j}\right] + \sum_j B_{\theta,kj} \mathbb{E}_{\theta,q}\left[\frac{\partial}{\partial\theta_i} u^I_{\theta,j}\right] + \mathbb{E}_{\theta,q}\left[\frac{\partial}{\partial\theta_i} a_{\theta,j}\right] \\ &= -\sum_j B_{\theta,kj} \mathbb{E}_{\theta,q}\left[u_{\theta,i} u^I_{\theta,j}\right] - \mathbb{E}_{\theta,q}\left[u_{\theta,i} a_{\theta,j}\right] = -\sum_j B_{\theta,kj} \mathbb{E}_{\theta,q}\left[u^I_{\theta,i} u^I_{\theta,j}\right]\end{aligned}$$

holds. In the above equalities, we use the formula

$$\mathbb{E}_{\theta,q}\left[\frac{\partial}{\partial\theta_i} f_\theta\right] = -\mathbb{E}_{\theta,q}\left[u_{\theta,i} f_\theta\right]$$

for $f_\theta \in T_{\theta,q}$ that is obtained by differentiating the identity $\mathbb{E}_{\theta,q}[f_\theta] = 0$. Clearly, $\boldsymbol{f}_\theta(x,e)$ provides the same estimator as $B_\theta^{-1}\boldsymbol{f}_\theta(x,e)$. See [19] for details of the standard form of Z-estimators.

Theorem 1. *Let us define the d by d matrix G_θ^I by*

$$G_\theta^I = \mathbb{E}[\boldsymbol{u}_\theta^I(\boldsymbol{u}_\theta^I)^T].$$

Then, for a fixed conditional probability $q(e|x)$, the asymptotic variance of any stochastic local Z-estimator $\widehat{\theta}$ satisfies the inequality

$$\lim_{m\to\infty} m \cdot V[\widehat{\theta}] \succeq (G_\theta^I)^{-1}$$

in the sense of the non-negative definiteness. The equality is attained by the Z-estimator using $\boldsymbol{u}_\theta^I$.

Proof. Let us compute each matrix in (9). According to the above argument, without loss of generality, we assume $\boldsymbol{f}_\theta = \boldsymbol{u}_\theta^I + \boldsymbol{a}_\theta$ for $\boldsymbol{a}_\theta \in (T_{\theta,q}^A)^d$. The matrix $\mathbb{E}_{\theta,q}[\boldsymbol{u}_\theta \boldsymbol{f}_\theta^T]$ is then expressed as

$$\mathbb{E}_{\theta,q}[\boldsymbol{u}_\theta \boldsymbol{f}_\theta^T] = \mathbb{E}_{\theta,q}[(\boldsymbol{u}_\theta^I + \mathbb{E}_{\theta,q}[\boldsymbol{u}_\theta|e])(\boldsymbol{u}_\theta^I + \boldsymbol{a}_\theta)^T] = \mathbb{E}_{\theta,q}[\boldsymbol{u}_\theta^I(\boldsymbol{u}_\theta^I)^T] = G_\theta^I$$

due to $\boldsymbol{u}_\theta^I \in (T_{\theta,q}^I)^d$, $\mathbb{E}[\boldsymbol{u}_\theta|e] \in (T_{\theta,q}^E)^d$ and $\boldsymbol{a}_\theta \in (T_{\theta,q}^A)^d$. Let us define $A = \mathbb{E}[\boldsymbol{a}_\theta \boldsymbol{a}_\theta^T]$. Then, we have

$$\mathbb{E}_{\theta,q}[\boldsymbol{f}_\theta \boldsymbol{f}_\theta^T] = G_\theta^I + A.$$

As a result, we obtain

$$\lim_{m\to\infty} m \cdot V[\widehat{\theta}] = (G_\theta^I)^{-1}(G_\theta^I + A)(G_\theta^I)^{-1} = (G_\theta^I)^{-1} + (G_\theta^I)^{-1}A(G_\theta^I)^{-1} \succeq (G_\theta^I)^{-1}.$$

When $\boldsymbol{f}_\theta = \boldsymbol{u}_\theta^I$, the matrix A becomes the null matrix and the minimum asymptotic variance is attained. □

The minimum variance of stochastic local Z-estimators is attained by the efficient score. This conclusion agrees to the result of the asymptotically efficient estimator in semi-parametric models including nuisance parameters [19,22].

Remark 1. Let us consider the relation between the stochastic pseudo-likelihood $\nabla \log p_\theta(x_k|x_{-k})$ and efficient score $\boldsymbol{u}_\theta^I(x, e_{x,k})$. Suppose that the neighborhood system N_x and the conditional distribution $q(e|x)$ on N_x are defined as shown in Example 3. Then, we have $\boldsymbol{u}_\theta^I(x, e_{x,k}) = \nabla \log p_\theta(x_k|x_{-k})$. Likewise, we find that the reduced-SCL, $\nabla \log p_\theta(x_{A_\ell}|x_{B_\ell})$, is equivalent with the efficient score under the setup in Example 4 when the index subset A_ℓ is defined as B_ℓ^c.

4.3. Monotonicity of Asymptotic Efficiency

As described in [23], for the composite likelihood estimator with the index pairs (A_ℓ, B_ℓ), $\ell = 1, \ldots, M$, it is widely believed that by increasing the size of A_ℓ (and correspondingly decreasing the size of $B_\ell = A_\ell^c$), one can capture more dependency relations in the model and increase the accuracy. For the stochastic local Z-estimators, we can obtain the exact relation between the neighborhood system and asymptotic efficiency.

Let us consider two stochastic local Z-estimators; one is defined by $q(e|x)$ on the neighborhood system $e \in N_x$ and the other is given by $q'(e'|x)$ on the neighborhood system $e' \in N_x'$. The efficient score are respectively written as $\boldsymbol{u}_\theta^I(x,e)$ for $q(e|x)$ and $\boldsymbol{u}_\theta'^I(x,e')$ for $q'(e|x)$. In addition, let us define $G_\theta^I = \mathbb{E}_{\theta,q}[\boldsymbol{u}_\theta^I(\boldsymbol{u}_\theta^I)^T]$ and $G_\theta'^I = \mathbb{E}_{\theta,q'}[\boldsymbol{u}_\theta'^I(\boldsymbol{u}_\theta'^I)^T]$.

Theorem 2. *Let $p(x,e,e')$ be the joint probability of $(x,e,e') \in \mathcal{X} \times N \times N'$ and suppose that probability functions, $q(e|x), q'(e'|x)$ and $p_\theta(x)$, are obtained from $p(x,e,e')$. We assume that*

$$\mathbb{E}[\mathbb{E}[\boldsymbol{u}_\theta|e]|e'] = \mathbb{E}[\boldsymbol{u}_\theta|e'] \tag{10}$$

holds under the probability distribution $p(x,e,e')$. Then, we have

$$(G_\theta^I)^{-1} \succeq (G_\theta'^I)^{-1}, \tag{11}$$

i.e., the efficiency bound of N_x' and $q'(x|e)$ is smaller than or equal to that of N_x and $q(x|e)$.

Proof. We use the basic formula of the conditional variance

$$\mathbb{V}[X] = \mathbb{V}[\mathbb{E}[X|Z]] + \mathbb{E}[\mathbb{V}[X|Z]] \succeq \mathbb{V}[\mathbb{E}[X|Z]] \tag{12}$$

for random variables X and Z. The above formula is applied to the score $\boldsymbol{u}_\theta(x)$ and the efficient score $\boldsymbol{u}_\theta^I(x,e)$. Note that $\mathbb{E}[\mathbb{V}[\boldsymbol{u}_\theta|e]] = \mathbb{E}[\boldsymbol{u}_\theta^I(\boldsymbol{u}_\theta^I)^T] = G_\theta^I$ holds. Then, we have

$$\begin{aligned}\mathbb{V}[\boldsymbol{u}_\theta] &= \mathbb{V}[\mathbb{E}[\boldsymbol{u}_\theta|e]] + \mathbb{E}[\mathbb{V}[\boldsymbol{u}_\theta|e]] = \mathbb{V}[\mathbb{E}[\boldsymbol{u}_\theta|e]] + \mathbb{E}[\boldsymbol{u}_\theta^I(\boldsymbol{u}_\theta^I)^T] \\ &= \mathbb{V}[\mathbb{E}[\boldsymbol{u}_\theta|e]] + G_\theta^I = \mathbb{V}[\mathbb{E}[\boldsymbol{u}_\theta|e']] + G_\theta'^I.\end{aligned}$$

The last equality comes from the fact that the score $\boldsymbol{u}_\theta(x)$ is common in both setups. Since the equality (10) holds, again the Formula (12) with $X = \mathbb{E}[\boldsymbol{u}_\theta|e]$ and $Z = e'$ yields

$$\mathbb{V}[\mathbb{E}[\boldsymbol{u}_\theta|e]] = \mathbb{V}[\mathbb{E}[\boldsymbol{u}_\theta|e']] + \mathbb{E}[\mathbb{V}[\mathbb{E}[\boldsymbol{u}_\theta|e]|e']] \succeq \mathbb{V}[\mathbb{E}[\boldsymbol{u}_\theta|e']].$$

Thus, we obtain

$$G_\theta^I = \mathbb{V}[\boldsymbol{u}_\theta] - \mathbb{V}[\mathbb{E}[\boldsymbol{u}_\theta|e]] \preceq \mathbb{V}[\boldsymbol{u}_\theta] - \mathbb{V}[\mathbb{E}[\boldsymbol{u}_\theta|e']] = G_\theta'^I.$$

As a result, we have (11). □

A similar inequality is derived in [24] for the mutual Fisher information. The mutual Fisher information is rather similar to $\mathbb{V}[\mathbb{E}[\boldsymbol{u}_\theta|e]]$ than G_θ^I. Theorem 13 of [24] corresponds to the one-dimensional version of the inequality $\mathbb{V}[\mathbb{E}[\boldsymbol{u}_\theta|e]] \succeq \mathbb{V}[\mathbb{E}[\boldsymbol{u}_\theta|e']]$.

Let us show an example that agrees to (10). We define two neighborhood systems $N = \{N_x | x \in \mathcal{X}\}$ and $N' = \{N_x' | x \in \mathcal{X}\}$ such that, for any $e \in N_x$, there exists $e' \in N_x'$ satisfying $e \subset e'$. For the joint probability $p(x,e,e')$, suppose that x and e' are conditionally independent given e and that the conditional probability $r'(e'|e)$ derived from $p(x,e,e')$ is equal to zero unless $e \subset e'$. Under these conditions, $q'(e'|x)$ derived from $p(x,e,e')$ takes 0 if $e' \notin N_x'$. The conditional independence assures that $p(x,e,e')$ is expressed as $p(x,e,e') = p_\theta(x)q(e|x)r'(e'|e) = p(x|e)r(e|e')q(e')$. Hence, the conditional probability $p(x|e')$ is expressed as $\sum_{e\in N} p(x|e)r(e|e')$. Thus, we obtain

$$\mathbb{E}[\mathbb{E}[\boldsymbol{u}_\theta|e]|e'] = \sum_{e\in N}\sum_{x\in\mathcal{X}} \boldsymbol{u}_\theta(x)p(x|e)r(e|e') = \sum_{x\in\mathcal{X}} \boldsymbol{u}_\theta(x) \sum_{e\in N} p(x|e)r(e|e') = \sum_{x\in\mathcal{X}} \boldsymbol{u}_\theta(x)p(x|e').$$

As a result, the better efficiency bound is obtained by the larger neighborhood. A similar result is presented in [25] for the composite likelihood estimators. The relation of the result in [25] and ours is explained in Section 5.3 of this paper.

Example 5. Let N_x be a neighborhood system at x endowed with the conditional distribution $q(e|x)$. Another neighborhood system is defined as $N_x' = \{\mathcal{X}\}$ for all x, and $q'(e'|x) = 1$ for $e' = \mathcal{X}$. Let us define $p(x,e,e') = p_\theta(x)q(e|x)$ for $e' = \mathcal{X}$ and otherwise $p(x,e,e') = 0$. Since e' always takes $\mathcal{X}$, x and e'

are conditionally independent given e. Thus, we have $G^I_\theta \preceq G'^I_\theta$. Indeed, G'^I_θ is the Fisher information matrix of the model $p_\theta(x)$.

We compare the stochastic pseudo-likelihood and reduced-SCL. Let $N_x = \{e_{x,k} \,|\, k = 1, \ldots, n\}$ be the neighborhood system defined in Example 3, and N be $\cup_{x\in\mathcal{X}} N_x$. The conditional distribution on N_x is given by $q(e_{x,k}|x) = q_k$, $k = 1, \ldots, n$. As shown in Remark 1, the corresponding efficient score is nothing but the stochastic pseudo-likelihood, i.e., $\boldsymbol{u}^I_\theta(x, e_{x,k}) = \nabla \log p_\theta(x_k|x_{-k})$. Let us define another neighborhood system N'_x in the same way as Example 4. For the subsets $B_\ell \subset \mathcal{X}$ and $A_\ell = B^c_\ell$, $\ell = 1, \ldots, M$, we define $e'_{x,\ell}$ as $\{y \in \mathcal{X} \,|\, y_{B_\ell} = x_{B_\ell}\}$ and $N'_x = \{e'_{x,\ell} \,|\, \ell = 1, \ldots, M\}$. Let N' be $\cup_{x\in\mathcal{X}} N'_x$. The conditional distribution on N'_x is given as $q'(e'_{x,\ell}|x) = q'_\ell$ for $\ell = 1, \ldots, M$. Then, the efficient score associated with N' and q' is equal to the reduced-SCL, i.e., $\boldsymbol{u}'^I_\theta(x, e'_{x,\ell}) = \nabla \log p_\theta(x_{A_\ell}|x_{B_\ell})$. As the direct conclusion of Theorem 2 and the above argument about the property of the conditional independence between x and $e' \in N'$ given $e \in N$, we obtain the following corollary.

Corollary 1. *We define N'_e for $e \in N$ by $N'_e = \{e' \in N' | e \subset e'\}$. Let $r'(e'|e)$ be a conditional probability on N'_e given $e \in N$, where $r'(e'|e) = 0$ is assumed for $e' \notin N'_e$. If the equality $q'_\ell = \sum_{k=1}^n q_k r'(e'_{x,\ell}|e_{x,k})$ holds, the reduced-SCL with N' and q' is more efficient than stochastic pseudo-likelihood with N and q.*

Example 6. Suppose that the size of N'_e is the same for all $e \in N$ and that the size of the set $\{e \in N | x \in e \subset e'\}$ is the same for any $x \in \mathcal{X}$ and $e' \in N'$ such that $x \in e'$. Let $q(e|x)$ (resp. $q'(e'|x)$) be the uniform distribution on N_x (resp. N'_x). Then, the reduced-SCL is more efficient than stochastic pseudo-likelihood. Indeed, the assumption ensures that the sum $\sum_{e\in N} q(e|x)r'(e'|e)$ does not depend on x and e'. Thus, the uniform distribution $q'(e'|x)$ meets the condition of the above corollary. For example, let $B_1, \ldots, B_M$ be all subsets of size $n-2$ in $\{1, \ldots, n\}$. Then, we have $M = n(n-1)/2$. The size of N'_e is $n-1$, and the size of $\{e \in N | x \in e \subset e'\}$ is equal to 2.

5. Local Z-Estimators and Efficiency Bounds

In this section, we define the local Z-estimator as the expectation of a stochastic local Z-estimator, and derive its efficiency bound.

5.1. Local Z-Estimators

Computationally tractable estimators such as pseudo-likelihood and composite likelihood are obtained by the expectation of an identification function in $T^L_{\theta,q}$. Let us define the *local Z-estimator* as the Z-estimator using

$$\bar{\boldsymbol{f}}_\theta(x) = \mathbb{E}_{\theta,q}[\boldsymbol{f}_\theta|x],$$

where $\boldsymbol{f}_\theta \in (T^L_{\theta,q})^d$. The conditional expectation given x is regarded as the projection onto the subspace $T^X_{\theta,q}$ which is defined as

$$T^X_{\theta,q} = \{f \in T_{\theta,q} \,|\, f(x, e) \text{ does not depend on } e\}.$$

Let Π_X be the projection operator onto $T^X_{\theta,q}$ and $\Pi^\perp_X$ be the one onto the orthogonal complement of $T^X_{\theta,q}$. Then, one can prove $\Pi_X[f] = \mathbb{E}[f|x]$ and $\Pi^\perp_X[f] = f - \mathbb{E}[f|x]$ for $f \in T_{\theta,q}$. When the number of elements in the neighborhood N_x is reasonable, the computation of the local Z-estimator is tractable.

Below, we show that some estimators are expressed as the local Z-estimator.

Example 7 (Pseudo-likelihood and composite likelihood)**.** In the setup of Example 3, the conditional expectation of the efficient score, $\mathbb{E}_{\theta,q}[\boldsymbol{u}^I_\theta|x]$, yields the pseudo-likelihood when $q(e|x)$ is the uniform distribution on N_x. In the setup of Example 4, let us assume $A_\ell = B^c_\ell$ and $q'(e'_{x,\ell}|x) = q'_\ell$. Then, the conditional expectation of the efficient score $\boldsymbol{u}^I_\theta(x, e'_{x,\ell}) = \nabla \log p_\theta(x_{A_\ell}|x_{B_\ell})$ yields

$$\mathbb{E}_{\theta,q'}[\boldsymbol{u}_\theta^I|x] = \sum_{\ell=1}^{M} q'_\ell \nabla \log p_\theta(x_{A_\ell}|x_{B_\ell}),$$

which is the general form of the composite likelihood in Example 2 with $\gamma_\ell = q'_\ell$.

5.2. Efficiency Bounds

We derive the efficiency bound of the local Z-estimator. Without loss of generality, the local Z-estimator $\bar{\boldsymbol{f}}_\theta(x) \in (T^X_{\theta,q})^d$ is represented as

$$\bar{\boldsymbol{f}}_\theta(x) = \mathbb{E}[\boldsymbol{f}_\theta|x], \quad \boldsymbol{f}_\theta = \boldsymbol{u}_\theta^I + \boldsymbol{a}_\theta \in (T^L_{\theta,q})^d, \ \ \boldsymbol{a}_\theta \in (T^A_{\theta,q})^d.$$

Under the model $p_\theta(x)$, we calculate the asymptotic variance (9) of the local Z-estimator $\widehat{\theta}$ using $\bar{\boldsymbol{f}}_\theta(x)$. The matrix $\mathbb{E}_{\theta,q}[\boldsymbol{u}_\theta \bar{\boldsymbol{f}}_\theta^T]$ in (9) is given as

$$\mathbb{E}_{\theta,q}[\boldsymbol{u}_\theta \bar{\boldsymbol{f}}_\theta^T] = \mathbb{E}_{\theta,q}[\boldsymbol{u}_\theta (\boldsymbol{u}_\theta^I + \boldsymbol{a}_\theta)^T] = \mathbb{E}[\boldsymbol{u}_\theta^I (\boldsymbol{u}_\theta^I)^T] = G_\theta^I.$$

Hence, we have

$$\lim_{m\to\infty} m \cdot \mathbb{V}[\widehat{\theta}] = (G_\theta^I)^{-1} \mathbb{E}_{\theta,q}[\bar{\boldsymbol{f}}_\theta \bar{\boldsymbol{f}}_\theta^T] (G_\theta^I)^{-1}.$$

Here, the expectation $\mathbb{E}_{\theta,q}[\bar{\boldsymbol{f}}_\theta \bar{\boldsymbol{f}}_\theta^T]$ can be written as the expectation under $p_\theta(x)$, i.e., $\mathbb{E}_\theta[\bar{\boldsymbol{f}}_\theta \bar{\boldsymbol{f}}_\theta^T]$, since $\boldsymbol{u}_\theta$ and $\bar{\boldsymbol{f}}_\theta$ depend only on x. The orthogonal decomposition $\boldsymbol{f}_\theta = \bar{\boldsymbol{f}}_\theta + \Pi_X^\perp[\boldsymbol{f}_\theta]$ leads to

$$\begin{aligned}(G_\theta^I)^{-1} \mathbb{E}_{\theta,q}[\boldsymbol{f}_\theta \boldsymbol{f}_\theta^T] (G_\theta^I)^{-1} &= (G_\theta^I)^{-1} \mathbb{E}_{\theta,q}[\bar{\boldsymbol{f}}_\theta \bar{\boldsymbol{f}}_\theta^T] (G_\theta^I)^{-1} + (G_\theta^I)^{-1} \mathbb{E}_{\theta,q}[\Pi_X^\perp[\boldsymbol{f}_\theta] \Pi_X^\perp[\boldsymbol{f}]_\theta^T] (G_\theta^I)^{-1} \\ &\succeq (G_\theta^I)^{-1} \mathbb{E}_{\theta,q}[\bar{\boldsymbol{f}}_\theta \bar{\boldsymbol{f}}_\theta^T] (G_\theta^I)^{-1},\end{aligned} \tag{13}$$

meaning that the asymptotic variance of the stochastic local Z-estimator using $\boldsymbol{f}_\theta(x, e)$ is larger then or equal to that of the local Z-estimator using $\bar{\boldsymbol{f}}_\theta(x)$.

We consider the optimal choice of $\boldsymbol{a}_\theta \in (T^A_{\theta,q})^d$ in $\bar{\boldsymbol{f}}_\theta(x) = \mathbb{E}_{\theta,q}[\boldsymbol{u}_\theta^I + \boldsymbol{a}_\theta|x]$. Let us define the subspace $T^{XA}_{\theta,q}$ as $\Pi_X T^A_{\theta,q} = \{\Pi_X[a] \,|\, a \in T^A_{\theta,q}\}$, and Π_{XA} be the projection operator onto $T^{XA}_{\theta,q}$. Then, we define $v^I_{\theta,j}(x) \in T^X_{\theta,q}$ as the projection of $u^I_{\theta,j}(x, e) \in T^I_{\theta,q}$ onto the orthogonal complement of $T^{XA}_{\theta,q}$ in $T^X_{\theta,q}$, i.e.,

$$v^I_{\theta,j} = (\Pi_X - \Pi_{XA})[u^I_{\theta,j}]$$

for $j = 1, \ldots, d$. In this paper, we call $\boldsymbol{v}^I_\theta = (v^I_{\theta,1}, \ldots, v^I_{\theta,d})^T$ the *local efficient score*.

Theorem 3. *Let us define d by d matrix H^I_θ as $\mathbb{E}_{\theta,q}[\boldsymbol{v}^I_\theta (\boldsymbol{v}^I_\theta)^T]$. Then, the efficiency bound of the local Z-estimator $\widehat{\theta}$ is given as*

$$\lim_{m\to\infty} m \cdot \mathbb{V}[\widehat{\theta}] \succeq (G_\theta^I)^{-1} H_\theta^I (G_\theta^I)^{-1}.$$

The equality is attained by the local Z-estimator using the local efficient score $\boldsymbol{v}^I_\theta = (\Pi_X - \Pi_{XA})[\boldsymbol{u}^I_\theta]$.

Proof. $\bar{\boldsymbol{f}}_\theta(x) = \mathbb{E}_{\theta,q}[\boldsymbol{u}^I_\theta + \boldsymbol{a}_\theta|x]$ has the orthogonal decomposition $\boldsymbol{v}^I_\theta + \boldsymbol{b}_\theta$, where $\boldsymbol{b}_\theta \in (T^{XA}_{\theta,q})^d$. Hence, we obtain $\mathbb{E}_{\theta,q}[\bar{\boldsymbol{f}}_\theta \bar{\boldsymbol{f}}_\theta^T] \succeq \mathbb{E}_\theta[\boldsymbol{v}^I_\theta (\boldsymbol{v}^I_\theta)^T] = H^I_\theta$ and

$$(G_\theta^I)^{-1} \mathbb{E}_{\theta,q}[\bar{\boldsymbol{f}}_\theta \bar{\boldsymbol{f}}_\theta^T] (G_\theta^I)^{-1} \succeq (G_\theta^I)^{-1} H_\theta^I (G_\theta^I)^{-1}.$$

The left-hand side of the above inequality is the asymptotic variance of the local Z-estimator. The equality is attained by the local Z-estimator using $\boldsymbol{v}^I_\theta$. □

We consider the relation between the local efficient score $v_\theta^I(x)$ and the score $u_\theta(x)$. We define $T_{\theta,q}^{ML}$ as the subspace spanned by the score $u_{\theta,j}(x), j = 1, \ldots, d$. For any $a \in T_{\theta,q}^A$, we have

$$\mathbb{E}_{\theta,q}[u_{\theta,j}\mathbb{E}_{\theta,q}[a|x]] = \mathbb{E}_{\theta,q}[u_{\theta,j}a] = 0, \quad j = 1, \ldots, d,$$

meaning that $T_{\theta,q}^{ML}$ and $T_{\theta,q}^{XA}$ are orthogonal to each other. Hence, $T_{\theta,q}^X$ is decomposed into

$$T_{\theta,q}^X = T_{\theta,q}^{XA} \oplus T_{\theta,q}^{ML} \oplus T_{\theta,q}^{XC},$$

where $T_{\theta,q}^{XC}$ is the orthogonal complement of $T_{\theta,q}^{XA} \oplus T_{\theta,q}^{ML}$ in $T_{\theta,q}^X$. Eventually, subspaces in $T_{\theta,q}$ satisfy the following relations,

$$T_{\theta,q} = T_{\theta,q}^E \oplus T_{\theta,q}^I \oplus T_{\theta,q}^A, \quad T_{\theta,q}^X = (\Pi_X T_{\theta,q}^A) \oplus T_{\theta,q}^{ML} \oplus T_{\theta,q}^{XC}.$$

Let us define $T_{\theta,q}^{XI}$ as the subspace spanned by the local efficient score $v_{\theta,j}^I(x)$, $j = 1, \ldots, d$. Under a mild assumption, $T_{\theta,q}^{XI}$ and $T_{\theta,q}^{ML}$ has the same dimension. Since $\boldsymbol{v}_\theta^I(x)$ is orthogonal to $\Pi_X T_{\theta,q}^A$, $T_{\theta,q}^{XI}$ is included in $T_{\theta,q}^{ML} \oplus T_{\theta,q}^{XC}$. Hence, $T_{\theta,q}^{XC}$ is interpreted as the subspace expressing the information loss caused by the localization of the score $\boldsymbol{u}_\theta$.

5.3. Relation to Existing Works

5.3.1. Comparison of Local Z-Estimators

We compare the asymptotic variances of two local Z-estimators that are connected to composite likelihoods.

One estimator is defined from the neighborhood system N which consists of the singleton $N_x = \{e_x\}$, $x \in \mathcal{X}$. Here, we assume that $e_x = e_{x'}$ holds for $x' \in e_x$ and $\cup_{x\in\mathcal{X}} e_x = \mathcal{X}$. Such a neighborhood system N is called the equivalence class [25]. An equivalence class corresponds to a partition of the sample space. The conditional probability $q(e|x)$ takes 1 for $e = e_x$ and 0 otherwise. Let $\boldsymbol{u}_\theta^I(x, e)$ be the efficient score defined from N and $q(e|x)$, and $\bar{\boldsymbol{u}}_\theta^I(x)$ be the local Z-estimator $\bar{\boldsymbol{u}}_\theta^I(x) = \mathbb{E}_{\theta,q}[\boldsymbol{u}_\theta^I|x]$.

Another localized estimator is defined from the neighborhood system N' which consists of $N'_x, x \in \mathcal{X}$, where N'_x is not necessarily a singleton. Suppose that $e_x \subset e'$ holds for any $e' \in N'_x$. The conditional probability $q'(e'|x)$ is defined as $q'(e'|x) = r'(e'|e_x)$, where $r'(e'|e_x)$ is a conditional probability of $e' \in N'_x$ given e_x. The corresponding efficient score is denoted as $\boldsymbol{u}_\theta'^I(x, e)$ and let us define $\bar{\boldsymbol{u}}_\theta'^I(x) = \mathbb{E}_{\theta,q'}[\boldsymbol{u}_\theta'^I|x]$ as the local Z-estimator associated with N' and $q'(e'|x)$.

From the definition, the joint probability $p_\theta(x)q(e|x)r'(e'|e)$ agrees to $q(e|x)$ an $q'(e'|x)$. Hence we see that x and e' are conditionally independent given e. Hence, Theorem 2 guarantees the inequality $(G_\theta^I)^{-1} \succeq (G_\theta'^I)^{-1}$.

The efficient score $\boldsymbol{u}_\theta^I(x, e)$ can take a non-zero value only when $e = e_x$. Hence, $\boldsymbol{u}_\theta^I(x, e)$ is regarded as the function of x, i.e, $\boldsymbol{u}_\theta^I(x, e) \in (T_{\theta,q}^X)^d$, and the asymptotic variance of the local Z-estimator obtained by $\bar{\boldsymbol{u}}_\theta^I(x) = \boldsymbol{u}_\theta^I(x, e_x)$ is $(G_\theta^I)^{-1}$. On the other hand, the asymptotic variance of the local Z-estimator derived from $\bar{\boldsymbol{u}}_\theta'^I(x)$ is less than or equal to $(G_\theta'^I)^{-1}$ due to (13). Therefore, $\bar{\boldsymbol{u}}_\theta'^I$ with N' and q' provides more efficient estimators than $\bar{\boldsymbol{u}}_\theta^I$ with N and q.

Liang and Jordan presented a similar result in [25]. In their setup, the larger neighborhood N'_x is a singleton $\{e'_x\}$ and the smaller one, N_x, can have multiple neighborhoods at each x. In such a case, the similar relation holds, i.e., the estimator with N' is more efficient. However, their approach is different from ours. In [25], the randomness is introduced over the patterns of the partition of $\mathcal{X}$. Moreover, their identification function corresponding to our $\bar{\boldsymbol{u}}_\theta^I(x)$ is decomposed into two terms; one is the term conditioned on the partition and the other is its orthogonal complement. On the other hand, our approach uses the decomposition of $\boldsymbol{u}_\theta^I(x, e)$ into $\bar{\boldsymbol{u}_\theta}^I(x)$ and its orthogonal complement.

In their analysis, the simplified expression of the asymptotic variance shown in (9) and the standard expression of the identification function, $\boldsymbol{f}(x,e) = \boldsymbol{u}^I_\theta(x,e) + \boldsymbol{a}(x,e)$, are not used. Hence, the evaluation of the asymptotic variance yields rather a complex dependency on the estimator. As a result, their approach does not show the efficiency bound, though the asymptotic variance of the composite likelihood for exponential families is presented under the misspecified setup.

5.3.2. Closed Exponential Families

The so-called closed exponential family has an interesting property from the viewpoint of localized estimators, as presented in [26]. Let $p_\theta(x) = \exp\{\theta^T t(x) - c(\theta)\}$ be the exponential family defined for $x = (x_1, \ldots, x_n) \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ with the parameter $\theta \in \Theta \subset \mathbb{R}^d$. The function $t(x) \in \mathbb{R}^d$ is referred to as the sufficient statistic. Given disjoint index subsets $A, B \subset \{1, \ldots, n\}$, let $t_B(x)$ be all elements of $t(x)$ that depend just on x_B, and $t_{A,B}(x)$ be the other elements. Hence, $t_B(x)$ is expressed as $t_B(x_B)$. The parameter θ is correspondingly decomposed into $\theta = (\theta_{A,B}, \theta_B)$. Thus, we have $\theta^T t(x) = \theta^T_{A,B} t_{A,B}(x) + \theta^T_B t_B(x_B)$. The exponential family $p_\theta(x)$ is called the closed exponential family, when the marginal distribution of x_B is expressed as the exponential family with the sufficient statistic $t_B(x_B)$.

We consider the composite likelihood of the closed exponential family. For the pairs of two disjoint index subsets, $\{A_\ell, B_\ell\}$, $\ell = 1, \ldots, M$, suppose that any element of $t(x)$ is included in $t_{A_\ell,B_\ell}(x)$ at least one ℓ. Then, the local Z-estimator using the composite likelihood $\sum_{\ell=1}^M \log p_\theta(x_{A_\ell}|x_{B_\ell})$ is identical to the MLE [26]. Hence, the composite likelihood of the closed exponential family attains the efficiency bound of the MLE.

For the general statistical model $p_\theta(x)$, let us restate the above result in terms of the tangent spaces in $T_{\theta,q}$. Let us decompose $p_\theta(x)$ into

$$p_\theta(x) = p(x_A|x_B;\theta)p(x_B;\theta).$$

We assume that for any index subset B, all elements of $\nabla \log p(x_B;\theta)$ are included in $T^{ML}_{\theta,q}$ that is spanned by the elements of $\boldsymbol{u}_\theta(x) = \nabla \log p_\theta(x)$. Then, $\nabla \log p(x_A|x_B;\theta)$ also lies in $(T^{ML}_{\theta,q})^d$. Thus, $\nabla \log p(x_{A_\ell}|x_{B_\ell};\theta)$ is expressed as $C_\ell \nabla \log p_\theta(x)$ using a d by d matrix C_ℓ. If $\sum_{\ell=1}^M C_\ell$ is invertible, the local Z-estimator obtained by $\sum_{\ell=1}^M \nabla \log p_\theta(x_{A_\ell}|x_{B_\ell})$ is identical to the MLE. In this case, $\Pi_X T^I_{\theta,q} = T^{ML}_{\theta,q}$, i.e., $T^{XI}_{\theta,q} = T^{ML}_{\theta,q}$ holds. Therefore, there is no information loss caused by the localization. The matrix C_ℓ for the closed exponential family is given as the projection matrix onto the subspace spanned by $t_{A_\ell,B_\ell}(x) - \mathbb{E}_\theta[t_{A_\ell,B_\ell}|x_B]$ that is included in $T^{ML}_{\theta,q}$. The above result implies that the tangent space $T^{XC}_{\theta,q}$ expressing the information loss will be related to the score of the marginal distribution, $\nabla \log p_\theta(x_B)$.

6. Conclusions

In this paper, some statistical properties of stochastic local Z-estimators and local Z-estimators are investigated. The class of local Z-estimators includes pseudo-likelihood and composite likelihood. For stochastic local Z-estimators, we established the exact relation between neighborhood systems and the efficiency bound under a simple and general condition. In addition, the efficiency bound of the local Z-estimators was presented.

Future works include the study of more general class of localized estimators. Indeed, local Z-estimators do not include the class of proper local scoring rules [7]. It is worthwhile to derive the efficiency bound for more general localized estimators. Exploring nice applications of the efficiency bound will be another interesting direction of our study. In our setup, the local efficient score expressed by the projection of the score attains the efficiency bound among local Z-estimators. An important problem is to develop a computationally tractable method to obtain the projection onto tangent subspaces.

Conflicts of Interest: The author declares no conflict of interest.

Appendix. Asymptotic Variance of Stochastic Local Z-Estimators

Given i.i.d. samples $(x_i, e_i), i = 1, \dots, m$ from $p_\theta(x)q(e|x)$, we estimate the parameter θ using the stochastic local Z-estimator obtained by

$$\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{f}(x_i, e_i; \widehat{\theta}) = \mathbf{0},$$

where the identification function satisfies $\boldsymbol{f}_\theta \in (T_{\theta,q})^d$ for any $\theta \in \Theta \subset \mathbb{R}^d$. The Taylor expansion around the true parameter θ yields

$$\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{f}(x_i, e_i; \theta) + \frac{1}{m}\sum_{i=1}^{m} \nabla \boldsymbol{f}(x_i, e_i; \theta)(\widehat{\theta} - \theta) + O(\|\widehat{\theta} - \theta\|^2) = \mathbf{0},$$

where the element $(\nabla \boldsymbol{f})_{ij}$ is given as $\frac{\partial f_i}{\partial \theta_j}$. As m tends to infinity, the asymptotic distribution of $\widehat{\theta}$ is given as the multivariate normal distribution,

$$\mathbb{E}_{\theta,q}[\nabla \boldsymbol{f}_\theta]\sqrt{m}(\widehat{\theta} - \theta) \sim N_d(\mathbf{0}, \mathbb{E}_{\theta,q}[\boldsymbol{f}_\theta \boldsymbol{f}_\theta^T]).$$

Since $\mathbb{E}_{\theta,q}[\boldsymbol{f}_\theta] = \mathbf{0}$ holds for any θ, the derivative $\nabla \mathbb{E}_{\theta,q}[\boldsymbol{f}_\theta]$ is the null matrix. This fact yields

$$\mathbb{E}_{\theta,q}[\nabla \boldsymbol{f}_\theta] = -\mathbb{E}_{\theta,q}[\boldsymbol{f}_\theta \nabla \log(p_\theta q)^T] = -\mathbb{E}_{\theta,q}[\boldsymbol{f}_\theta \boldsymbol{u}_\theta^T].$$

Hence, the asymptotic distribution of $\sqrt{m}(\widehat{\theta} - \theta)$ is the d-dimensional normal distribution with mean $\mathbf{0}$ and variance $\mathbb{E}_{\theta,q}[\boldsymbol{f}_\theta \boldsymbol{u}_\theta^T]^{-1}\mathbb{E}_{\theta,q}[\boldsymbol{f}_\theta \boldsymbol{f}_\theta^T]\mathbb{E}_{\theta,q}[\boldsymbol{u}_\theta \boldsymbol{f}_\theta^T]^{-1}$.

References

1. Van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 2000.
2. Besag, J. Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B* **1974**, *36*, 192–236.
3. Lindsay, B.G. Composite likelihood methods. *Contemp. Math.* **1988**, *80*, 220–239.
4. Varin, C.; Reid, N.; Firth, D. An overview of composite likelihood methods. *Stat. Sin.* **2011**, *21*, 5–42.
5. Hyväinen, A. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Trans. Neural Netw.* **2007**, *18*, 1529–1531.
6. Hyvärinen, A. Some extensions of score matching. *Comput. Stat. Data Anal.* **2007**, *51*, 2499–2512.
7. Dawid, A.P.; Lauritzen, S.; Parry, M. Proper local scoring rules on discrete sample spaces. *Ann. Stat.* **2012**, *40*, 593–608.
8. Kanamori, T.; Takenouchi, T. Graph-Based Composite Local Bregman Divergences on Discrete Sample Spaces. 2016, arXiv:1604.06568.
9. Lafferty, J.D.; McCallum, A.; Pereira, F.C.N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, San Francisco, CA, USA, 28 June–1 July 2001; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2001; pp. 282–289.
10. Ackley, H.; Hinton, E.; Sejnowski, J. A learning algorithm for boltzmann machines. *Cognit. Sci.* **1985**, *9*, 147–169.
11. Smolensky, P. Information Processing in Dynamical Systems: Foundations of Harmony Theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*; MIT Press: Cambridge, MA, USA, 1986; pp. 194–281.
12. Welling, M.; Rosen-Zvi, M.; Hinton, G.E. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems 17*; Saul, L.K., Weiss, Y., Bottou, L., Eds.; MIT Press: Cambridge, MA, USA, 2005; pp. 1481–1488.
13. Ising, E. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik* **1925**, *31*, 253–258. (In German)

14. Marlin, B.; de Freitas, N. Asymptotic efficiency of deterministic estimators for discrete energy-based models: Ratio matching and pseudolikelihood. In *Uncertainty in Artificial Intelligence (UAI)*; AUAI Press: Corvallis, OR, USA, 2011.
15. Gneiting, T. Making and evaluating point forecasts. *J. Am. Stat. Assoc.* **2011**, *106*, 746–762.
16. Steinwart, I.; Pasin, C.; Williamson, R.C.; Zhang, S. Elicitation and identification of properties. In Proceedings of the 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, 13–15 June 2014; pp. 482–526.
17. Hyväinen, A. Consistency of pseudolikelihood estimation of fully visible boltzmann machines. *Neural Comput.* **2006**, *18*, 2283–2292.
18. Dawid, A.; Musio, M. Estimation of spatial processes using local scoring rules. *AStA Adv. Stat. Anal.* **2013**, *97*, 173–179.
19. Amari, S.; Kawanabe, M. Information geometry of estimating functions in semi-parametric statistical models. *Bernoulli* **1997**, *3*, 29–54.
20. Dillon, J.V.; Lebanon, G. Stochastic composite likelihood. *J. Mach. Learn. Res.* **2010**, *11*, 2597–2633.
21. Cichocki, A.; Amari, S. Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, 1532–1568.
22. Bickel, P.J.; Klaassen, C.A.J.; Ritov, Y.; Wellner, J.A. *Efficient and Adaptive Estimation for Semiparametric Models*; Springer-Verlag: New York, NY, USA, 1998.
23. Asuncion, A.U.; Liu, Q.; Ihler, A.T.; Smyth, P. Learning with blocks: Composite likelihood and contrastive divergence. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; Volume 9, pp. 33–40.
24. Zegers, P. Fisher information properties. *Entropy* **2015**, *17*, 4918–4939.
25. Liang, P.; Jordan, M.I. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In Proceedings of the 25th International Conference on Machine Learning, ICML '08, Helsinki, Finland, 5–9 July 2008; ACM: New York, NY, USA, 2008; pp. 584–591.
26. Mardia, K.V.; Kent, J.T.; Hughes, G.; Taylor, C.C. Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika* **2009**, *96*, 975–982.

© 2016 by the author; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).