Open Journal of Statistics, 2014, 4, 188-197
Published Online April 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.43018

Theoretical Properties of Composite Likelihoods

Xiaogang Wang, Yuehua Wu

Department of Mathematics and Statistics, York University, Toronto, Canada
Email: stevenw@mathstat.yorku.ca, wuyh@mathstat.yorku.ca

Received 7 January 2014; revised 7 February 2014; accepted 15 February 2014

Abstract

The general functional form of composite likelihoods is derived by minimizing the Kullback-Leibler distance under structural constraints associated with low dimensional densities. Connections with the *I*-projection and the *maximum entropy distributions* are shown. Asymptotic properties of composite likelihood inference under the proposed information-theoretical framework are established.

Keywords

Composite Likelihood; I-Divergence; Information Theory; Likelihood Weights; Maximum Entropy Distribution

1. Introduction

The composite likelihood has been increasingly used when the full likelihood is computationally intractable or difficult to specify due to either high dimensionality or complex dependence structures. Consider a random vector $\boldsymbol{X}$ with probability density $f(\boldsymbol{x};\boldsymbol{\theta})$, where $\boldsymbol{x}=(x_1,\cdots,x_p)^{\mathrm{T}}\in R^p$ and $\boldsymbol{\theta}\in R^d$. Denote the component likelihoods by $L_k(\boldsymbol{\theta};\boldsymbol{x})$, where $k=1,2,\cdots,K$, and the composite likelihood proposed in [1] is defined by

$$L_C(\boldsymbol{\theta};\boldsymbol{x})=\prod_{k=1}^{K}L_k(\boldsymbol{\theta};\boldsymbol{x})^{\lambda_k},$$

where λ_k's are non-negative weights to be chosen.

As discussed in [2], there are two general types of composite likelihood: marginal and conditional composite likelihood. The simplest composite likelihood is the one constructed under the independence assumption:

$$L_{\text{indep}}(\boldsymbol{\theta};\boldsymbol{x})=\prod_{r=1}^{p}f(x_r;\boldsymbol{\theta}).$$

How to cite this paper: Wang, X.G. and Wu, Y. (2014) Theoretical Properties of Composite Likelihoods. *Open Journal of Statistics*, **4**, 188-197. http://dx.doi.org/10.4236/ojs.2014.43018

If the inferential interest is also on parameters prescribing a dependence structure, a pairwise composite likelihood [2] [3] is defined as the following:

$$L_{\text{pairwise-indep}}(\boldsymbol{\theta};\boldsymbol{x})=\prod_{r=1}^{p}\prod_{s=1}^{p} f(x_r,x_s;\boldsymbol{\theta}).$$

Conditional composite likelihood [4] [5] can be constructed by multiplying all pairwise conditional densities:

$$L_{\text{conditional}}(\boldsymbol{\theta};\boldsymbol{x})=\prod_{r=1}^{p}\prod_{s=1}^{p} f(x_r|x_s;\boldsymbol{\theta}).$$

There are other important variations and applications of the composite likelihoods designed for various inferential purposes such as composite likelihood BIC for model selection in high-dimensional data in [6]. Detailed discussions and review of composite likelihoods were provided in [2].

Since there are various composite likelihoods with different functional forms, it might be desirable to consider a unifying theme based on information-theoretic justifications. Under an information-theoretic framework, composite likelihoods can then be viewed as a class of inferential functions based on optimal probability density under structural constraints imposed on low dimensional densities when the complete joint density is either unknown or untractable. We show that the optimal densities associated with the composite likelihood are also connected with the *I*-projection density well-known in probability theory and the maximum entropy distributions in information theory. Although likelihood weights are employed in the original formulation of composite likelihood in [1], equal weights are often adopted due to convenience. We show that adaptive likelihood weights can indeed improve the performance of composite likelihood inference using equal weights.

This paper is organized as follows. In Section 2, we derive the composite likelihood as the optimal inferential device by minimizing the relative entropy or Kullbak-Leilber distance under structural constraints. Asymptotic properties are established in Section 3. Discussions are given in Section 4.

2. Derivation of Composite Likelihood with Weights

2.1. *I*-Projection and Maximum Entropy Distribution

Suppose that $g(\boldsymbol{x})$ and $f(\boldsymbol{x})$ are generalized densities of a dominated set of probability measures on the measurable space $(\Omega,\mathcal{F})$. The relative entropy is defined as

$$I(g,f)=\int g(\boldsymbol{x})\log\left[\frac{g(\boldsymbol{x})}{f(\boldsymbol{x})}\right]\mathrm{d}\lambda(\boldsymbol{x}).$$

The relative entropy is widely used in information theory and also known as *I*-divergence in probability. In [7], Cover and Thomas provide an excellent account on its properties and applications in information theory and coding theory. As demonstrated in [8], the relative entropy can play an important role in statistical inference. The relative entropy is also called *I*-divergence and its geometric properties are studied in [9]. Although the relative entropy or *I*-divergence is not a metric and in general does not define a topology, Csiszár in [9] shows that certain analogies exist between properties of probability distributions and Euclidean geometry, where *I*-divergence plays the role of squared distance. It is a measure of discrepancy between the probability densities *g* and *f*.

For any probability density function (pdf) f_0, Csiszár in [9] defines an *I*-sphere centered around f_0 with a radius ρ as the following:

$$S(f_0,\rho)=\{g:I(g,f_0)<\rho,0<\rho\le\infty\},$$

where *g* is a probability density function.

In statistical inference, the pdf f_0 is the model of choice when the true pdf is unknown. In high dimensional or complex cases, it is high unlikely that the assumed model f_0 is correct. When no other information on the dependence structure is available, the best model might be the one based on the independent assumption.

When significant characteristics associated with the low dimensional projections of the joint probability density function, it is then desirable to incorporate this information formally into the statistical inference. To improve the chosen model, one might utilize constraints associated with known features under an information theoretic framework to be described in the following. As in [8], one might consider minimizing $I(g, f_0)$ with respect to g subject to

$$\int \boldsymbol{t}(\boldsymbol{x}) g(\boldsymbol{x}) \mathrm{d}\lambda(\boldsymbol{x}) = \boldsymbol{d}, \tag{1.1}$$

where $\boldsymbol{d}$ is a constant vector and $\boldsymbol{t}(\boldsymbol{x})$ a measurable multivariate statistic.

If $\mathcal{E}$ is a convex set of pdf intersecting $S(f_0, \rho)$, an optimal pdf g^* satisfying

$$I(g^*, f_0) = \min_{g \in \mathcal{E}} I(g, f_0) \tag{1.2}$$

is defined as the *I*-projection of f_0 on $\mathcal{E}$ in [9]. If such a projection exists, the convexity of $\mathcal{E}$ guarantees its uniqueness since $I(g, f_0)$ is strictly convex in *g*.

The following theorem follows immediately from the above theorem in [9].

Theorem 1. *Given pdf's* $f_0, f_1, f_2, \cdots, f_m$, *define*

$$\mathcal{H} = \bigcap_{i=1}^{m} \mathcal{E}_i^{(I)}(a_i),$$

where, for $i = 1, 2, \cdots, m$,

$$\mathcal{E}_i^{(I)}(a_i) = \left\{ g : \int g(\boldsymbol{x}) \log\left[f_i(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x} = a_i \right\}.$$

Then the optimal probability density function (the I-projection of f_0*)* $g_1^* \in \mathcal{H} \cap S(f_0, \infty)$ *takes the form*

$$g_1^*(\boldsymbol{x}) = D_1 f_0(\boldsymbol{x}) \left[f_1(\boldsymbol{x}) \right]^{\alpha_1} \left[f_2(\boldsymbol{x}) \right]^{\alpha_2} \cdots \left[f_m(\boldsymbol{x}) \right]^{\alpha_m},$$

where $D_1 = D_1(a_1, a_2, \cdots, a_m)$ *is the normalizing constant.*

Similar to the *I*-projection, the *maximum entropy distribution* is also an optimal density under constraints. It is also known as the Maxwell-Boltzmann distribution, the optimal probability density function under temperature constraints. Consider the following maximization problem:

$$\max_g H(g) = \max_g \left(-\int g(\boldsymbol{x}) \log\left[g(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x} \right)$$

in which $g(\boldsymbol{x})$ satisfying

$$\begin{aligned} &\int g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1, g(\boldsymbol{x}) \geq 0, \\ &\int g(\boldsymbol{x}) r_i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = b_i, i = 0, 1, \cdots, m. \end{aligned} \tag{1.3}$$

By applying the maximum entropy theorem in [7] with the constraints set as the logarithm of certain density functions, we then have the following result.

Theorem 2. *Let* $f_0, f_1, \cdots, f_m$ *be a set of probability density functions. If we set* $r_i = \log(f_i)$, $i = 0, 1, 2, \cdots, m$, *then there exists one unique maximum entropy density function that takes the form*:

$$g^*(\boldsymbol{x}) = D_3 \left[f_0(\boldsymbol{x}) \right]^{\beta_0} \prod_{i=1}^{m} \left[f_i(\boldsymbol{x}) \right]^{\beta_i},$$

where $D_3(\beta_0, \beta_1, \cdots, \beta_m)$ *is the normalizing constant.*

It is clear that the *I*-projection and the maximum entropy distribution could belong to the same functional class when a set of pdf's are used to formulate the constraints.

2.2. Derivation of Composite Likelihood Using Pseudo-Metric

If we consider the functional space of all probability density functions satisfying certain conditions and adopt the relative entropy as a pseudo-metric, then a more natural view of point is to seek an optimal density minimizing the relative entropy with constraints characterized by the pseudo distance between the optimal density and a collection of candidate models, $f_0, f_1, f_2, \cdots, f_m$.

In the context of composite likelihoods, the statistical model f_0 is the joint statistical model assumed while other pdf's are low dimensional densities to be used to complete the construction of a refined model which may or may not coincides with f_0. For example, one could assume a statistical model under an independence structure, *i.e.*, $f_0 = \prod_{i=1}^{m} f_i$ where $f_1, f_2, \cdots, f_m$ are low dimensional probability density functions. The composite likelihood framework, however, is capable of going beyond this often over-simplified model.

To ensure that the optimal density reflects some known key characteristics in the low dimensional densities of the true pdf, one can apply the idea of *I*-projection or maximum entropy distribution by considering the following minimization problem:

$$\int g(\boldsymbol{x}) \log\left[\frac{g(\boldsymbol{x})}{f_i(\boldsymbol{x})}\right] \mathrm{d}\boldsymbol{x} = c_i[f], i = 1, 2, \cdots, m, \tag{1.4}$$

where $c_1[f], c_2[f], \cdots, c_m[f]$ are functions of the true joint pdf *f*. The constraints employed here are different and more natural than those in the *I*-projection and maximum entropy formulation. In the original setup of the *I*-projection and maximum entropy distribution, the constraints are expectations of some certain statistics. The theorems of *I* projection and maximum entropy, however, are no longer applicable as the current set of constraints involves $\log g(\boldsymbol{x})$.

We now present our main theorem of this section.

Theorem 3. *Given probability density functions* $f_0, f_1, \cdots, f_m$ *, define*

$$\mathcal{H}^{(2)} = \bigcap_{i=1}^{m} \mathcal{E}_i^{(2)}(c_i[f]),$$

where, for $i = 1, \cdots, m$ *,*

$$\mathcal{E}_i^{(2)}(c_i) = \left\{ g : \int g(\boldsymbol{x}) \log\left[\frac{g(\boldsymbol{x})}{f_i(\boldsymbol{x})}\right] \mathrm{d}\boldsymbol{x} = c_i[f] \right\}.$$

Then the optimal probability density function satisfying

$$I(g^*, f_0) = \min_{g \in \mathcal{H}^{(2)} \cap S(f_0, \infty)} I(g, f_0)$$

takes the form

$$g^*(\boldsymbol{x}) = D f_0(\boldsymbol{x})^{\lambda_0} f_1(\boldsymbol{x})^{\lambda_1} f_2(\boldsymbol{x})^{\lambda_2} \cdots f_m(\boldsymbol{x})^{\lambda_m},$$

where $D(\lambda_1, \lambda_2, \cdots, \lambda_m)$ *is a normalizing constant and* $\sum_{i=0}^{m} \lambda_i = 1$.

The assertion of this theorem implies that the constraints in the original *I*-projection can be further generalized such that they are also a functionals of the probability density we seek as well. It can also be seen that $\mathcal{E}_i^{(2)}(c_i[f]) = S(g, c_i[f])$, the sphere in the functional space of all probability functions as in the context of *I*-projection.

The optimal pdf under the current constraints belongs to the following functional class:

$$\mathcal{G} = \left\{ g(\boldsymbol{x}) \middle| g(\boldsymbol{x}) \propto \prod_{i=0}^{m} f_i(\boldsymbol{x})^{\lambda_i} \right\}, \tag{1.5}$$

where $f_1, f_2, \cdots, f_m$ are low dimensional density functions.

We now consider four special cases:

1) (INDEPENDENT CASE) For example, if we assume that $f_i(\boldsymbol{x}_j) = \prod_{i=1}^{p} f_{[i]}(x_{ij})$, the marginals. Note that we use $f_{[i]}$ to denote the marginals in order to distinguish them from the probability density f_i used in the construction. If we set $f_0(\boldsymbol{x}) = \prod_{j=1}^{n} \prod_{i=1}^{p} f_{[i]}(x_{ij})$, it then implies that the constraints, which are based on the marginals only, do not bring in any additional structural information than f_0. Therefore, it follows that the optimal functional density is of the form

$$f^{*}(\boldsymbol{x}) \propto \prod_{j=1}^{n}\prod_{i=1}^{p} f_{[i]}(x_{ij}).$$

if all the weights equal to 1.

2) (CORRELATION CASE) If the constraints are defined by $f_i = f_{[ij]}(x_i, x_j)$ and $f_0(\boldsymbol{x}) = \prod_{j=1}^{n}\prod_{i=1}^{p} f_{[i]}(x_{ij})$, it then follows that

$$f^{*}(\boldsymbol{x}) \propto \prod_{j=1}^{n}\prod_{i=1}^{p} f_{[i]}(x_{ij}) \prod_{j=1}^{n}\prod_{i=1}^{p}\prod_{t=1}^{p} f_{[ij]}(x_{ij}, x_{it})^{\alpha_{ij}}.$$

The optimal density is then constructed by the marginals and all pairwise bivariate densities. A simplified form is given by

$$f^{*}(\boldsymbol{x}) \propto \prod_{j=1}^{n}\prod_{i=1}^{p} f_{[i]}(x_{ij}) \prod_{j=1}^{n}\prod_{i=1}^{p}\prod_{t=1}^{p} f_{[ij]}(x_{ij}, x_{it})^{\alpha}.$$

if $\alpha_{ij} = \alpha$.

3) (CONDITIONAL CASE) If the constraints are defined by $f(x_{i1}) f(x_{i2}, x_{i3} | x_{i1})$, we can then derive the conditional composite likelihood.

4) (SPATIAL AND TEMPORAL CASE) The weights might be most appropriate for the spatial or temporal settings. Consider $y_{ijts} = x_{ij} - x_{ts}$ for some given *t* and *i*. The composite likelihood can also be derived if the Jacobian for transformation is ignored due to its complexity. This would allow spatial and temporal correlation structure to be incorporated.

3. Asymptotic Properties of Composite Likelihood

In this section, we establish the asymptotic properties associated with the composite likelihood inference under the proposed information-theoretic framework. The consistency of the estimators is proved by following the argument in [10].

For clear presentation, we first define the following notations:

- Denote the true density function by $f_{\text{true}}(\cdot)$. Let $\{f_1(\cdot,\boldsymbol{\theta}),\cdots, f_K(\cdot,\boldsymbol{\theta})\}$ be the set of density function components under consideration.
- Denote $\boldsymbol{\lambda} = (\lambda_1,\cdots,\lambda_K)$. The set of probability density functions

$$\mathcal{F} = \left\{ c_{\boldsymbol{\theta},\boldsymbol{\lambda}} \prod_{k=1}^{K} \left[f_k(\cdot,\boldsymbol{\theta}) \right]^{\lambda_k}, \lambda_k \geq 0, 1 \leq k \leq K \right\}$$

with $\boldsymbol{\theta} \in \Omega \subset R^d$ and $\boldsymbol{\lambda} \in \Xi \subset (0,\infty)^K$ may not contain the true density function $f_{\text{true}}(\cdot)$. Put $\boldsymbol{\vartheta} = \left(\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\lambda}^{\mathrm{T}}\right)^{\mathrm{T}}$ and $c_{\boldsymbol{\theta},\boldsymbol{\lambda}} \prod_{k=1}^{K} \left[f_k(\cdot,\boldsymbol{\theta}) \right]^{\lambda_k} \triangleq f(\cdot,\boldsymbol{\vartheta})$.

- Let $\mathfrak{D}(\cdot,\cdot)$ be the distance function defined over the space of all density functions. Assume that there is a unique $f(\cdot,\boldsymbol{\vartheta}^*) \in \mathcal{F}$ such that $\mathfrak{D}\left(f_{\text{true}}, f(\cdot,\boldsymbol{\vartheta}^*)\right) = \min_{f\in\mathcal{F}} \mathfrak{D}(f_{\text{true}}, f)$. We further assume that $\mathfrak{D}\left(f_{\text{true}}, f(\cdot,\boldsymbol{\vartheta}^*)\right) > 0$ if $f_{\text{true}} \notin \mathcal{F}$. For demonstration, $\mathfrak{D}(\cdot,\cdot)$ is chosen as the K-L divergence in this paper.
- Let $\hat{\boldsymbol{\vartheta}}$ be the estimate of $\boldsymbol{\vartheta}$ such that

$$\hat{\boldsymbol{\vartheta}} = \arg\sup_{\boldsymbol{\vartheta}} \prod_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{\vartheta}).$$

Define

$$f_k(\boldsymbol{x},\boldsymbol{\theta},\varrho) = \sup\nolimits_{\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\|\leq\varrho} f_k(\boldsymbol{x},\tilde{\boldsymbol{\theta}}), k = 1,\cdots,K;$$

$$f_k^{*}(\boldsymbol{x},\boldsymbol{\theta},\varrho) = f_k(\boldsymbol{x},\boldsymbol{\theta},\varrho) \vee 1, k = 1,\cdots,K;$$

$$g_k(\boldsymbol{x},\tau) = \sup\nolimits_{\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\|>\tau} f_k(\boldsymbol{x},\tilde{\boldsymbol{\theta}}), k = 1,\cdots,K;$$

$$g_k^*(\boldsymbol{x},\tau)=g_k(\boldsymbol{x},\tau)\vee 1, k=1,\cdots,K;$$

$$f(\boldsymbol{x},\boldsymbol{\vartheta},\rho)=\sup\nolimits_{\|\hat{\vartheta}-\vartheta\|\le\rho} f\left(x,\hat{\boldsymbol{\vartheta}}\right);$$

$$f^*(\boldsymbol{x},\boldsymbol{\vartheta},\rho)= f(\boldsymbol{x},\boldsymbol{\vartheta},\rho)\vee 1;$$

$$g(\boldsymbol{x},\kappa)=\sup\nolimits_{\|\boldsymbol{\vartheta}\|>\kappa} f(\boldsymbol{x},\boldsymbol{\vartheta});$$

$$g^*(\boldsymbol{x},\kappa)=g(\boldsymbol{x},\kappa)\vee 1.$$

We make the following assumptions.

Assumption 1. $f_1,\cdots,f_K$ are measurable, and linearly independent in probability.

Assumption 2. For $k=1,\cdots,K$, $E\left[\log f_k^*(\boldsymbol{X},\boldsymbol{\theta},\varrho)\right]<\infty$ for sufficiently small ϱ and $E\left[\log g_k^*(\boldsymbol{X},\tau)\right]<\infty$ for sufficiently large τ.

Assumption 3. If $\boldsymbol{\theta}^{(j)}\to\boldsymbol{\theta}$ as $j\to\infty$, then for $k=1,\cdots,K$,

$$\lim_{i\to\infty} f_k\left(\boldsymbol{x},\boldsymbol{\theta}^{(j)}\right)= f_k(\boldsymbol{x},\boldsymbol{\theta}),\text{a.s.}$$

Assumption 4. If $\lim_{j\to\infty}\left\|\boldsymbol{\theta}^{(j)}\right\|=\infty$, then for $k=1,\cdots,K$,

$$\lim_{j\to\infty} f_k\left(\boldsymbol{x},\boldsymbol{\theta}^{(j)}\right)=0,\text{a.s.}$$

Assumption 5. $P\left[f(\cdot,\boldsymbol{\vartheta}_1)\ne f(\cdot,\boldsymbol{\vartheta}_2)\right]>0$ if $\boldsymbol{\vartheta}_1\ne\boldsymbol{\vartheta}_2$.

Assumption 6. Ξ is a closed set.

Assumption 7. Ω is a closed set.

We first give four lemmas in the following before we present the theorems regarding the limiting behavior of the weighted composite likelihood estimators.

Lemma 1. *The following hold true*:

(*L*1) *Under Assumption* 1, $f(\boldsymbol{x},\boldsymbol{\vartheta})$ *is measurable, and hence for any* $\varrho>0$, $f(\boldsymbol{x},\boldsymbol{\vartheta},\varrho)$ *is measurable.*

(*L*2) *Under Assumption* 2, $E\left\{\log\left[f^*(\boldsymbol{x},\boldsymbol{\vartheta},\varrho)\right]\right\}$ *is finite for sufficiently small* κ *and* $E\left\{\log\left[g^*(\boldsymbol{x},\kappa)\right]\right\}$ *is finite for sufficiently large* κ.

(*L*3) *Assume that Assumption* 3 *holds. If* $\boldsymbol{\vartheta}^{(j)}\to\boldsymbol{\vartheta}$ *as* $j\to\infty$, *then*

$$\lim_{i\to\infty} f\left(\boldsymbol{x},\boldsymbol{\vartheta}^{(j)}\right)= f(\boldsymbol{x},\boldsymbol{\vartheta}),\text{a.s.}$$

(*L*4) *Assume that Assumptions* 4 *and* 7 *holds. If* $\lim_{j\to\infty}\left\|\boldsymbol{\vartheta}^{(j)}\right\|=\infty$, *then*

$$\lim_{j\to\infty} f\left(\boldsymbol{x},\boldsymbol{\vartheta}^{(j)}\right)=0,\text{a.s.}$$

Lemma 2. *Assume that Assumptions* 1, 2, 6 *hold. For any* $\boldsymbol{\vartheta}\ne\boldsymbol{\vartheta}^*$,

$$E\left\{\log\left[f(\boldsymbol{x},\boldsymbol{\vartheta})\right]\right\}<E\left\{\log\left[f\left(\boldsymbol{x},\boldsymbol{\vartheta}^*\right)\right]\right\}.$$

Lemma 3. *Assume that Assumptions* 1 - 3 *hold. Then*

$$\lim_{\rho\to0} E\left\{\log\left[f(\boldsymbol{X},\boldsymbol{\vartheta},\rho)\right]\right\}=E\left\{\log\left[f(\boldsymbol{X},\boldsymbol{\vartheta})\right]\right\}.$$

Lemma 4. *Assume that Assumptions* 1, 2, 4, 7 *hold. Then*

$$\lim_{\kappa\to\infty} E\left\{\log\left[g(\boldsymbol{X},\kappa)\right]\right\}=-\infty.$$

The four theorems describing the limiting behavior of the weighted composite likelihood estimators are given below.

Theorem 4. *Assume that Assumptions* 1 - 6 *hold. Let* ϖ *be any closed subset of* $\Omega\times\Xi$ *that does not contain* $\boldsymbol{\vartheta}^*$. *Then*

$$P\left\{\lim_{n\to\infty}\frac{\sup_{\vartheta\in\varpi}\prod_{i=1}^{n} f\left(\boldsymbol{x}_i,\vartheta\right)}{\prod_{i=1}^{n} f\left(\boldsymbol{x}_i,\vartheta^*\right)}=0\right\}=1. \tag{1.6}$$

Theorem 5. *Assume that Assumptions* 1 - 7 *hold. Let* $\tilde{\vartheta}_n$ *be a function of the random samples* $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$ *such that*

$$\frac{\prod_{i=1}^{n} f\left(\boldsymbol{x}_i,\tilde{\vartheta}_n\right)}{\prod_{i=1}^{n} f\left(\boldsymbol{x}_i,\vartheta^*\right)}\geq\delta>0$$

for any n and for all observations. Then

$$P\left(\lim_{n\to\infty}\tilde{\vartheta}_n=\vartheta^*\right)=1.$$

Theorem 6. *Assume that Assumptions* 1 - 7 *hold. Then* $\hat{\vartheta}\to\vartheta^*$ *, a.s.*

Remark 1. Note that in the proof of Theorem 4, the strong law of large numbers is used. If we prove it using the method given in [11], the consistency of ϑ may be extended to a large class of dependent observations.

Remark 2. For simple presentation, we have assumed that $\left\{f_1,\cdots,f_k\right\}$ are parametric. This restriction is not necessary.

In the following we assume that λ is a constant vector. For easy presentation, define $\breve{f}(\cdot,\boldsymbol{\theta})=\prod_{k=1}^{K}\left[f_k(\cdot,\boldsymbol{\theta})\right]^{\lambda_k}$. Let θ be a solution of the following equations:

$$\sum_{k=1}^{K}\lambda_k\frac{\partial\log f_k\left(\boldsymbol{x},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}=\boldsymbol{0}.$$

For convenience, denote

$$\left.\frac{\partial l\left(\boldsymbol{x},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}=\frac{\partial\log l\left(\boldsymbol{x},\boldsymbol{\theta}_0\right)}{\partial\boldsymbol{\theta}};$$

and

$$\left.\frac{\partial^2 l\left(\boldsymbol{x},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}=\frac{\partial^2 l\left(\boldsymbol{x},\boldsymbol{\theta}_0\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}},$$

for a twice differentiable function $l\left(\boldsymbol{x},\boldsymbol{\theta}\right)$. To investigate the limiting distribution of the composite likelihood estimator, we make the following three more assumptions.

Assumption 8. For each $k\in\left\{1,\cdots,K\right\}$, $f_k\left(\boldsymbol{x},\boldsymbol{\theta}\right)$ is twice continuously differentiable in $\boldsymbol{\theta}$, and satisfies

$$E\left[\frac{\partial\psi_k\left(\boldsymbol{x},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right]=\frac{\partial E\left[\psi_k\left(\boldsymbol{x},\boldsymbol{\theta}\right)\right]}{\partial\boldsymbol{\theta}},$$

where $\psi_k\left(\boldsymbol{x},\boldsymbol{\theta}\right)=f_k\left(\boldsymbol{x},\boldsymbol{\theta}\right)$ and $=\partial f_k\left(\boldsymbol{x},\boldsymbol{\theta}\right)/\partial\boldsymbol{\theta}$.

Assumption 9. $E\left\{\partial\log f_k\left(\boldsymbol{x},\boldsymbol{\theta}\right)/\partial\boldsymbol{\theta}\left[\partial\log f_k\left(\boldsymbol{x},\boldsymbol{\theta}\right)/\partial\boldsymbol{\theta}\right]^{\mathrm{T}}\right\}$ is positive definite, for $k=1,\cdots,K$.

Assumption 10. There exist a positive number $\tau_{\boldsymbol{\theta}}$ and a positive function $\zeta\left(\boldsymbol{x},\boldsymbol{\theta}\right)$ such that $E\left[\zeta\left(\boldsymbol{X},\boldsymbol{\theta}\right)\right]<\infty$ and

$$\sup_{\left\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\right\|<\tau_{\boldsymbol{\theta}}}\left\|\frac{\partial^2\log f_k\left(\boldsymbol{x},\tilde{\boldsymbol{\theta}}\right)}{\partial\tilde{\boldsymbol{\theta}}\partial\tilde{\boldsymbol{\theta}}^{\mathrm{T}}}\right\|\leq\zeta\left(\boldsymbol{x},\boldsymbol{\theta}\right)$$

for all $\boldsymbol{x}$ in the range of $\boldsymbol{X}_1$.

Define

$$H\left(\boldsymbol{\theta}\right)=\sum_{k=1}^{K}\lambda_k E\left\{\frac{\partial\log f_k\left(\boldsymbol{X},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\left[\frac{\partial\log f_k\left(\boldsymbol{X},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right]^{\mathrm{T}}\right\},$$

and

$$G(\boldsymbol{\theta}) = \sum_{k_1=1}^{K}\sum_{k_2=1}^{K} \lambda_{k_1}\lambda_{k_2} \times E\left\{\frac{\partial \log f_{k_1}(\boldsymbol{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\left[\frac{\partial \log f_{k_2}(\boldsymbol{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^{\mathrm{T}}\right\}$$

We have the following theorem.

Theorem 7. *Assume that Assumptions* 1 - 10 *hold. Then*

$$\sqrt{n}\left(\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) \to N_d\left\{\mathbf{0}, \left[H\left(\boldsymbol{\theta}^*\right)\right]^{-1} G\left(\boldsymbol{\theta}^*\right)\left[H\left(\boldsymbol{\theta}^*\right)\right]^{-1}\right\}.$$

Remark 3. In light of [12], the assumptions 1 - 8 made in Theorem 7 may be replaced by the assumptions similar to those assumed in Theorem 4.17 of Shao (2003).

Remark 4. Let $\hat{\boldsymbol{\theta}}$ be the solution of

$$n\frac{\partial \log\left(c_{\boldsymbol{\theta},\lambda}\right)}{\partial \boldsymbol{\theta}} + \sum_{k=1}^{K} \lambda_k \frac{\partial \log f_k(\boldsymbol{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

By modifying the proof of Theorem 7, $\hat{\boldsymbol{\theta}}$ can also be shown to be asymptotically normal distributed.

4. Concluding Remarks

The proposed information-theoretic framework provides theoretical justifications for the use of composite likelihood. It also serves as a unifying theme for various seemingly different composite likelihoods and connects them with *I*-projection and maximum entropy distribution. Significant characteristics of low dimensional models are incorporated into the constraints associated with component likelihoods. Asymptotic properties established in this article could be useful for further theoretical analysis of the properties of the composite likelihoods. The findings presented in this article will lead to more in-depth investigations on the theoretical properties of composite likelihoods and establish some possible connections with information theory.

References

[1] Lindsay, B. (1988) Composite Likelihood Methods. *Contemporary Mathematics*, **80**, 221-239. http://dx.doi.org/10.1090/conm/080/999014

[2] Varin, C., Reid, N. and Firth, D. (2011) An Overview of Composite Likelihood Methods. *Statistica Sinica*, **21**, 5-42.

[3] Cox, D. and Reid, N. (2011) An Note on Pseudo-Likelihood Constructed from Marginal Densities. *Biometrika*, **91**, 729-737. http://dx.doi.org/10.1093/biomet/91.3.729

[4] Mollenberghs, G. and Verbeke, G. (2005) Models for Discrete Longitudinal Data. Springer, Inc., New York.

[5] Mardia, K.V., Kent, J.T., Hughes, G. and Taylor, C.C. (2009) Maximum Likelihood Estimation Using Composite Likelihoods for Closed Exponential Families. *Biometrika*, **96**, 975-982. http://dx.doi.org/10.1093/biomet/asp056

[6] Gao, X. and Song, P.X. (2010) Composite Likelihood Bayesian Information Criteria for Model Selection in High-Dimensional Data. *Journal of the American Statistical Association*, **105**, 1531-1540. http://dx.doi.org/10.1198/jasa.2010.tm09414

[7] Cover, T.M. and Thomas, J.A. (2006) Elements of Information Theory. John Wiley & Sons, Inc., Hoboken.

[8] Kullback, S. (1959) Information Theory and Statistics. Dove Publications, Inc., New York.

[9] Csiszár, I. (1975) *I*-Divergence Geometry of Probability Distributions as Minimization Problems. *Annals of Probability*, **3**, 146-158. http://dx.doi.org/10.1214/aop/1176996454

[10] Wald, A. (1949) Note on the Consistency of the Maximum Likelihood Estimate. *Annals of Mathematical Statistics*, **20**, 595. http://dx.doi.org/10.1214/aoms/1177729952

[11] Wolfowitz, J. (1949) On Wald's Proof of the Consistency of the Maximum Likelihood Estimate. *Annals of Mathematical Statistics*, **20**, 601-602. http://dx.doi.org/10.1214/aoms/1177729953

[12] Shao, J. (2003) Mathematical Statistics. 2nd Edition, Springer, Inc., New York. http://dx.doi.org/10.1007/b97553

Appendix

Proof of Theorem 1: Let $\psi_i(\boldsymbol{x}) = \log[f_i(\boldsymbol{x})], i = 1, \cdots, m.$ The *I*-projection is of the form

$$g^*(\boldsymbol{x}) = c_1 \exp\left(\sum_{i=1}^{m} \alpha_i \log[f_i(\boldsymbol{x})]\right) f_0(\boldsymbol{x}) = c_1 f_0(\boldsymbol{x}) \prod_{i=1}^{m} [f_i(\boldsymbol{x})]^{\alpha_i} = c_1 f_0(\boldsymbol{x}) \prod_{i=1}^{m} [f_i(\boldsymbol{x})]^{\alpha_i}.$$

This completes the proof. ◊

Proof of Theorem 3: By the Lagrange method, we seek to minimize the following objective function

$$W(g) = \int g(\boldsymbol{x}) \log\left[\frac{g(\boldsymbol{x})}{f_0(\boldsymbol{x})}\right] \mathrm{d}\boldsymbol{x} + \sum_{i=1}^{m} \ell_i \left(\int g(\boldsymbol{x}) \log\left[\frac{g(\boldsymbol{x})}{f_i(\boldsymbol{x})}\right] \mathrm{d}\boldsymbol{x} - c_i\right) + \ell_0 \left(\int g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - 1\right),$$

where $\ell_0, \ell_1, \cdots, \ell_m$ are Lagrange multipliers.

The objective function can then be rearranged so that

$$W(g) = \int U(g(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} - \ell_0 - \sum_{i=1}^{m} \ell_i c_i,$$

where

$$U(g(\boldsymbol{x})) = g(\boldsymbol{x}) \log\left[\frac{g(\boldsymbol{x})}{f_0(\boldsymbol{x})}\right] + \sum_{i=1} \ell_i g(\boldsymbol{x}) \log\left[\frac{g(\boldsymbol{x})}{f_i(\boldsymbol{x})}\right] + \ell_0 g(\boldsymbol{x}).$$

Since $U(g)$ is not a function of g', the first order derivative of *g*, the Euler-Lagrange equation is then given by

$$\frac{\partial U(g)}{\partial g} = 0,$$

where the derivative is taken with respect to *g*.

Thus, we have

$$\left(1 + \sum_{i=1}^{m} \ell_i\right) \{1 + \log[g(\boldsymbol{x})]\} = \log[f_0(\boldsymbol{x})] + \sum_{i=1}^{m} \cdots \ell_i \log f_i(\boldsymbol{x}) - \ell_0.$$

It then follows that the optimal density function takes the form

$$g_2^*(\boldsymbol{x}) = \mathrm{e}^{-\frac{\ell_0 + s_\ell}{s_\ell}} [f_0(\boldsymbol{x})]^{1/s_\ell} \prod_{i=1}^{m} [f_i(\boldsymbol{x})]^{\ell_i/s_\ell},$$

where $s_\ell = 1 + \sum_{i=1}^{n} \ell_i$. ◊

Proof of Lemma 2: In view of the definition of ϑ^*, the properties of K-L divergence and Lemma 1, Lemma 2 can be proved by following the proof of Lemma 1 of Wald (1949) ◊.

Proof of Lemma 3: By Lemma 1, Lemma 3 can be proved by following the proof of Lemma 2 of Wald (1949). ◊

Proof of Lemma 4: By applying Lemma 1, Lemma 4 can be proved by following the proof of Lemma 3 of Wald (1949). ◊

Proof of Theorem 4: By Lemmas 2 and 4, we can find a positive number κ_0 such that

$$E\left[\log g^*(\boldsymbol{X}, \kappa_0)\right] < E\left[\log g^*(\boldsymbol{X}, \vartheta^*)\right]. \tag{1.7}$$

Let ϖ_1 be the subset of ϖ consisting of all points ϑ of $\Omega \times \Xi$ for which $\|\vartheta\| \le \kappa_0$. By Lemmas 2 - 3, for each point $\vartheta \in \varpi_1$, there is a $\rho_\vartheta > 0$ such that

$$E\{\log[f(\boldsymbol{X}, \vartheta, \rho_\vartheta)]\} < E\{\log[f(\boldsymbol{X}, \vartheta^*)]\}. \tag{1.8}$$

Since ϖ_1 is a closed set, there exists a finite number of points $\vartheta_1, \cdots, \vartheta_h$ in ϖ_1 such that

$$\bigcup_{j=1}^{h}\mathcal{O}\left(\boldsymbol{\vartheta}_j,\rho_{\boldsymbol{\vartheta}_j}\right)\supset\varpi_1,$$

where $\mathcal{O}\left(\boldsymbol{\vartheta}_j,\rho_{\boldsymbol{\vartheta}_j}\right)$ denotes the open sphere with center $\boldsymbol{\vartheta}_j$ and radius $\rho_{\boldsymbol{\vartheta}_j}$. Thus,

$$0\le\sup_{\boldsymbol{\vartheta}\in\varpi}\prod_{i=1}^{n} f\left(\boldsymbol{x}_i,\boldsymbol{\vartheta}\right)\le\sup_{\boldsymbol{\vartheta}\in\varpi_1}\prod_{i=1}^{n} f\left(\boldsymbol{x}_i,\boldsymbol{\vartheta}\right)+\sup_{\|\boldsymbol{\vartheta}\|>\kappa_0}\prod_{i=1}^{n} f\left(\boldsymbol{x}_i,\boldsymbol{\vartheta}\right)\le\sum_{j=1}^{h}\prod_{i=1}^{n} f\left(\boldsymbol{x}_i,\boldsymbol{\vartheta}_j,\rho_{\boldsymbol{\vartheta}_j}\right)+\prod_{i=1}^{n} g\left(\boldsymbol{x}_i,r_0\right).$$

In light of (1.7)-(1.8), we have

$$\frac{1}{n}\sum_{i=1}^{n}\left[\log\left(f\left(\boldsymbol{X}_i,\boldsymbol{\vartheta}_j,\rho_{\boldsymbol{\vartheta}_j}\right)\right)-\log\left(f\left(\boldsymbol{X}_i,\boldsymbol{\vartheta}^*\right)\right)\right]\to E\left\{\log\left[f\left(\boldsymbol{X},\boldsymbol{\vartheta}_j,\rho_{\boldsymbol{\vartheta}_j}\right)\right]\right\}-E\left\{\log\left[f\left(\boldsymbol{X},\boldsymbol{\vartheta}^*\right)\right]\right\}<0,\text{a.s.}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\left[\log\left(g\left(\boldsymbol{X}_i,\kappa_0\right)\right)-\log\left(f\left(\boldsymbol{X}_i,\boldsymbol{\vartheta}^*\right)\right)\right]\to E\left\{\log\left[g\left(\boldsymbol{X},\boldsymbol{\vartheta}_j,\rho_{\boldsymbol{\vartheta}_j}\right)\right]\right\}-E\left\{\log\left[f\left(\boldsymbol{X},\boldsymbol{\vartheta}^*\right)\right]\right\}<0,\text{a.s.}$$

Therefore,

$$P\left\{\lim_{n\to\infty}\sum_{i=1}^{n}\left[\log\left(f\left(\boldsymbol{X}_i,\boldsymbol{\vartheta}_j,\rho_{\boldsymbol{\vartheta}_j}\right)\right)-\log\left(f\left(\boldsymbol{X}_i,\boldsymbol{\vartheta}^*\right)\right)\right]=-\infty\right\}=1,$$

$$P\left\{\lim_{n\to\infty}\sum_{i=1}^{n}\left[\log\left(g\left(\boldsymbol{X}_i,\kappa_0\right)\right)-\log\left(f\left(\boldsymbol{X}_i,\boldsymbol{\vartheta}^*\right)\right)\right]=-\infty\right\}=1,$$

which jointly with (1.9) implies (1.6). ◊

Proof of Theorem 5: For any $\varepsilon>0$, if a subsequence $\left\{\tilde{\boldsymbol{\vartheta}}_{n_m}\right\}$ of $\left\{\tilde{\boldsymbol{\vartheta}}_n\right\}$ that has a limit $\tilde{\boldsymbol{\vartheta}}$ such that $\left\|\tilde{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}^*\right\|>\varepsilon$, then for infinitely many n,

$$\sup_{\left\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^*\right\|\ge\varepsilon}\prod_{i=1}^{n} f\left(\boldsymbol{x}_i,\boldsymbol{\vartheta}\right)>\prod_{i=1}^{n} f\left(\boldsymbol{x}_i,\tilde{\boldsymbol{\vartheta}}_n\right).$$

Hence, for infinitely many *n*,

$$\frac{\sup_{\left\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^*\right\|\ge\varepsilon}\prod_{i=1}^{n} f\left(\boldsymbol{x}_i,\boldsymbol{\vartheta}\right)}{\prod_{i=1}^{n} f\left(\boldsymbol{x}_i,\boldsymbol{\vartheta}^*\right)}\ge\delta>0.$$

By Theorem 4, this event has zero probability. Thus all limit points $\tilde{\boldsymbol{\vartheta}}$ of $\left\{\tilde{\boldsymbol{\vartheta}}_n\right\}$ satisfy the inequality $\left\|\tilde{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}^*\right\|\le\varepsilon$ with probability one, which concludes the theorem. ◊

Proof of Theorem 7: By following the proof of Theorem 4.17 of Shao (2003), it can be shown that

$$\frac{\partial\log f\left(\boldsymbol{x},\breve{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}}=\frac{\partial\log f\left(\boldsymbol{x},\boldsymbol{\theta}^*\right)}{\partial\boldsymbol{\theta}}-H\left(\boldsymbol{\theta}^*\right)\left(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)+o_p\left(\left\|H\left(\boldsymbol{\theta}^*\right)\left(\boldsymbol{\theta}-\boldsymbol{\theta}^*\right)\right\|\right).$$

Hence,

$$\frac{\partial\log f\left(\boldsymbol{x},\boldsymbol{\theta}^*\right)}{\partial\boldsymbol{\theta}}=H\left(\boldsymbol{\theta}^*\right)\left(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)+o_p\left(\left\|H\left(\boldsymbol{\theta}^*\right)\left(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right\|\right),$$

which, jointly with Slutsky's theorem and the central limit theorem, concludes the proof of the theorem. ◊