Title	Peter Hall's Contribution to Empirical Likelihood
Manuscript ID	SS-2017-0059.R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0059
Complete List of Authors	Jinyuan Chang, Jianjun Guo and Cheng Yong Tang
Corresponding Author	Jinyuan Chang
E-mail	changjinyuan@swufe.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Peter Hall's Contribution to Empirical Likelihood

Jinyuan Chang[†], Jianjun Guo[†*], Cheng Yong Tang[‡]
[†]Southwestern University of Finance and Economics, and [‡]Temple University

Abstract

We deeply mourn the loss of Peter Hall. Peter was the premier mathematical statistician of his era. His work illuminated many aspects of statistical thought. While his body of work on bootstrap and nonparametric smoothing is widely known and appreciated, less well known is his work in many other areas. In this article, we review Peter's contribution to empirical likelihood (EL). Peter has done fundamental work on studying the coverage accuracy of confidence regions constructed with EL.

1 Introduction

Empirical likelihood (EL) amounts to computing the profile likelihood of a general multinomial distribution which has its atoms at data points. A version of this technique dates back at least to Thomas and Grunkemeier (1975), in the context of estimating survival probabilities. Since introduced in the seminal papers of Owen (1988, 1990), EL has received broad attentions from various areas, and a rich literature has been developed on approaches constructed based on it. We quote a part of Peter's review from Prof. Owen's website that "It is a unique practical tool, and it enjoys important, and growing, connections to many areas of statistics, from the Kaplan-Meier estimator to the bootstrap and beyond." EL is nonparametric in the sense that less restrictive distributional assumptions are required. Also, EL is similar to the bootstrap, which is another nonparametric area that Peter has great influence, in many aspects as a nonparametric device for statistical inferences including hypothesis testing and confidence set estimation. As the most remarkable advantage, without stringent distributional assumption, EL is a more robust counterpart of the conventional parametric likelihood yet sharing two main merits – Wilks' phenomenon and Bartlett correction – of the conventional likelihood.

Hall and La Scala (1990) provided a first review of the EL methodology and algorithms. Hall and La Scala (1990) elaborates on four main advantages of EL over the bootstrap:

(i) Confidence regions constructed by EL are data driven without any explicit or implicit requirement on their shape. Thus the confidence regions are oriented by observed data, and

[*]Corresponding author. Tel: +86 28 87092418, Fax: +86 28 87092418, E-mail: guojj@swufe.edu.cn

tend to be concentrated in places where the empirical density of the parameter estimator is greatest.

(ii) EL regions are Bartlett correctable. That is, the order of the coverage error can be reduced from n^{-1} to n^{-2} with a simple correction for the mean EL ratio statistics, where n is the sample size.

(iii) No estimation of scale or skewness is required when applying EL for statistical inferences. Indeed, thanks to its self-Studentized property, EL does not even require construction of a pivotal statistic.

(iv) EL regions are range preserving and transformation respecting. That is, the EL region for the function $\mathbf{g}(\boldsymbol{\theta})$ of parameter $\boldsymbol{\theta}$ equals the set by obtained by applying $\mathbf{g}$ on the EL region for $\boldsymbol{\theta}$.

The monograph of Owen (2001) supplies a comprehensive text on methodological, theoretical, and computational aspects of EL. Later, Chen and Van Keilegom (2009) reviewed EL in the framework of regressions covering various parametric, nonparametric, and semiparametric models. Recently, the monograph of Zhou (2016) elaborates on EL and survival analysis.

As one of the most prominent researcher in nonparametric statistics, Peter's contribution to EL is highly influential, especially from theoretical and technical aspects that are fundamental for studying the properties of EL. For example, the Edgeworth expansion based analysis in DiCiccio, Hall and Romano (1991) pioneered the studies of the Bartlett correctibility of EL. A closer examination of the details reveals striking similarities between analyses in EL and the bootstrap especially when studying the coverage accuracies. Since early '90s, the literature studying EL has grown rapidly with applications in numerous areas. In this article, we review Peter's contribution to EL in Section 2, and some discussion of the current challenges and development of EL in Section 3. Our personal reminisces are given in Section 4.

2 Peter's contributions to EL

Peter published 6 papers on EL in the '90s. Three of them were on *The Annals of Statistics*, and the other three were published on *Biometrika*, *International Statistical Review* and *Journal of Computational and Graphical Statistics*, respectively. The main contributions of Peter's these papers are reviewed here.

2.1 Comparison between parametric likelihood and EL

Peter's first paper on EL is the one coauthored with Thomas J. DiCicco and Joseph P. Romano published on *Biometrika* (DiCiccio, Hall and Romano, 1989). In the paper, they comprehensibly examined the difference between the parametric likelihood and EL functions, or surfaces in the context that the functional of interest is a smooth function of some mean vector. They found that though the two surfaces in general need not to agree with each other in the first order, the two surfaces are quite close when data follow some exponential family models.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sample of d-dimensional random vectors with $\boldsymbol{\mu}_0 = \mathbb{E}(\mathbf{X}_i) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{X}_i) = \mathbf{I}_d$. For any $\boldsymbol{\mu} \in \mathbb{R}^d$, the EL for $\boldsymbol{\mu}$ is defined as $L_n(\boldsymbol{\mu}) = \sup\{\prod_{i=1}^n w_i : w_i \geq 0,\ \sum_{i=1}^n w_i = 1,\ \sum_{i=1}^n w_i \mathbf{X}_i = \boldsymbol{\mu}\}$. Write $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,d})^{\mathrm{T}}$ and define $\bar{\mathbf{X}}_n = n^{-1}\sum_{i=1}^n \mathbf{X}_i = (\bar{X}_{n,1}, \ldots, \bar{X}_{n,d})^{\mathrm{T}}$. Denote by $|\cdot|_2$ the L_2-norm of a vector. For any $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^{\mathrm{T}}$ such that $|\boldsymbol{\mu}|_2 \leq Cn^{-1/2}$ for some arbitrarily large but fixed constant C, if $\mathbb{E}(|\mathbf{X}_i|_2^6) < \infty$, they showed that the empirical log-likelihood ratio $\ell_E(\boldsymbol{\mu}) = -2\log\{L_n(\boldsymbol{\mu})/L_n(\bar{\mathbf{X}}_n)\}$ can be expressed as

$$\ell_E(\boldsymbol{\mu}) = n(\bar{\mathbf{X}}_n - \boldsymbol{\mu})^{\mathrm{T}}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) + \frac{2n}{3}\sum_{j,k,m=1}^{d} \alpha_{jkm}(\bar{X}_{n,j} - \mu_j)(\bar{X}_{n,k} - \mu_k)(\bar{X}_{n,m} - \mu_m)$$
$$- n\sum_{j,k=1}^{d} \Delta_{jk}(\bar{X}_{n,j} - \mu_j)(\bar{X}_{n,k} - \mu_k) + R_{n,1},$$

where $\alpha_{jkm} = \mathbb{E}(X_{i,j}X_{i,k}X_{i,k})$, $\Delta_{jk} = n^{-1}\sum_{i=1}^n X_{i,j}X_{i,k} - \mathbb{I}(j=k)$, and $R_{n,1}$ is $O\{n^{-1}(\log\log n)^2\}$ almost surely and $O(n^{-1})$ in probability. As for the parametric likelihood function, they considered $\mathbf{X}_i$ in the exponential family. Suppose that an s-dimensional random vector has a density function $f_{\boldsymbol{\lambda}}(\mathbf{y}) = \exp\{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{U}(\mathbf{y}) - \psi(\boldsymbol{\lambda})\}f_0(\mathbf{y})$, where $\mathbf{U}(\cdot)$ is an d-dimensional function of s variables, $\boldsymbol{\lambda}$ is an unknown r-dimensional parameter, and $\psi(\boldsymbol{\lambda}) = \log[\int \exp\{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{U}(\mathbf{y})\}f_0(\mathbf{y})d\mathbf{y}]$. Put $\mathbf{X} = \mathbf{U}(\mathbf{Y})$ with $\mathbb{E}_{\boldsymbol{\lambda}_0}(\mathbf{X}) = \mathbf{0}$ and $\mathrm{Var}_{\boldsymbol{\lambda}_0}(\mathbf{X}) = \mathbf{I}_r$, where $\boldsymbol{\lambda}_0$ is the true value of $\boldsymbol{\lambda}$. Given $\boldsymbol{\lambda}$, define $\boldsymbol{\mu} = \nabla_{\boldsymbol{\lambda}}\psi(\boldsymbol{\lambda})$. Then the parametric log-likelihood ratio $\ell_P(\boldsymbol{\mu})$ can be expressed as

$$\ell_P(\boldsymbol{\mu}) = n(\bar{\mathbf{X}}_n - \boldsymbol{\mu})^{\mathrm{T}}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) + \frac{2n}{3}\sum_{j,k,m=1}^{d} \alpha_{jkm}(\bar{X}_{n,j} - \mu_j)(\bar{X}_{n,k} - \mu_k)(\bar{X}_{n,m} - \mu_m)$$
$$- n\sum_{j,k,m=1}^{d} \alpha_{jkm}(\bar{X}_{n,j} - \mu_j)(\bar{X}_{n,k} - \mu_k)\bar{X}_{n,m} + R_{n,2},$$

where $R_{n,2}$ is $O\{n^{-1}(\log\log n)^2\}$ almost surely and is $O(n^{-1})$ in probability. Indeed,

$$\begin{aligned}&\ell_E(\boldsymbol{\mu}) - \ell_P(\boldsymbol{\mu})\\ &= n\sum_{j,k,m=1}^{r} \alpha_{jkm}(\bar{X}_{n,j} - \mu_j)(\bar{X}_{n,k} - \mu_k)\bar{X}_{n,m} - n\sum_{j,k=1}^{r} \Delta_{jk}(\bar{X}_{n,j} - \mu_j)(\bar{X}_{n,k} - \mu_k) + R_n,\end{aligned}$$

where R_n is $O\{n^{-1}(\log\log n)^2\}$ almost surely and is $O(n^{-1})$ in probability. This result implies that $\ell_E(\boldsymbol{\mu})$ and $\ell_P(\boldsymbol{\mu})$ differ in a term of precise order $n^{-1/2}$. Nevertheless, based on such expressions of $\ell_E(\boldsymbol{\mu}) - \ell_P(\boldsymbol{\mu})$ and applying the Edgeworth expansion, one can show that the distributions of $\ell_E(\boldsymbol{\mu})$ and $\ell_P(\boldsymbol{\mu})$ differ only up to the order $O(n^{-1})$, i.e., $\mathbb{P}\{\ell_E(\boldsymbol{\mu}) \leq z\} - \mathbb{P}\{\ell_P(\boldsymbol{\mu}) \leq z\} = O(n^{-1})$ for any given $z > 0$. This investigation deepens the understanding of EL, additional to the fact that $\ell_E(\mu)$ is asymptotically χ^2 distributed.

2.2 EL is Bartlett-correctable

Peter's second work on EL should be the one published on *The Annals of Statistics* in 1991 coauthored with Thomas J. DiCicco and Joseph P. Romano (DiCiccio, Hall and Romano, 1991), although he had two papers published respectively on *The Annals of Statistics* and *International Statistical Review* before it. DiCiccio, Hall and Romano (1991) is an influential paper on the coverage accuracy of the EL ratios. In the context that the functional of interest is a smooth function of vector means, the results in DiCiccio, Hall and Romano (1991) substantially strengthen the EL as tool for constructing confidence regions.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ denote a sample from an unknown d-variate distribution F_0 having mean $\boldsymbol{\mu}_0 = \mathbb{E}(\mathbf{X}_i)$ and nonsingular covariance matrix $\boldsymbol{\Sigma}_0 = \mathrm{Var}(\mathbf{X}_i)$. The parameter of interest is $\boldsymbol{\theta}_0 = \mathbf{K}(\boldsymbol{\mu}_0)$ for some given q-dimensional function $\mathbf{K}(\cdot) = \{K_1(\cdot), \ldots, K_q(\cdot)\}^{\mathrm{T}}$ with $q \le d$. In such a context, the EL for $\boldsymbol{\theta}$ is given by $L_n(\boldsymbol{\theta}) = \sup\{\prod_{i=1}^n w_i : w_i \ge 0,\ \sum_{i=1}^n w_i = 1,\ \mathbf{K}(\sum_{i=1}^n w_i \mathbf{X}_i) = \boldsymbol{\theta}\}$ and the empirical log-likelihood function $\ell_E(\boldsymbol{\theta}) = -2\log[L_n(\boldsymbol{\theta})/L_n\{\mathbf{K}(\bar{\mathbf{X}}_n)\}]$. The Wilks' theorem says that $\ell_E(\boldsymbol{\theta}_0) \to_d \chi_q^2$, as $n \to \infty$. More precisely, $\mathbb{P}\{\ell_E(\boldsymbol{\theta}_0) \le z\} = \mathbb{P}(\chi_q^2 \le z) + O(n^{-1})$ for any fixed $z > 0$. DiCiccio, Hall and Romano (1991) showed that EL admits a Bartlett correction. That is,

$$\mathbb{P}\big[\ell_E(\boldsymbol{\theta}_0)\{\mathbb{E}(n\mathbf{R}^{\mathrm{T}}\mathbf{R})/q\}^{-1} \le z\big] = \mathbb{P}(\chi_q^2 \le z) + O(n^{-2}),$$

where $\mathbf{R}$ is a q-dimensional vector, defined in Section 4 of DiCiccio, Hall and Romano (1991), such that $\ell_E(\boldsymbol{\theta}_0) = n\mathbf{R}^{\mathrm{T}}\mathbf{R} + O_p(n^{-3/2})$. Moreover, the ratio $\{\mathbb{E}(n\mathbf{R}^{\mathrm{T}}\mathbf{R})/q\}^{-1}$ has a simple expansion, $\{\mathbb{E}(n\mathbf{R}^{\mathrm{T}}\mathbf{R})/q\}^{-1} = 1 - an^{-1} + O(n^{-2})$, where a is a fixed constant. Replacing a by an estimate $\widehat{a}$, it holds that

$$\mathbb{P}\big[\ell_E(\boldsymbol{\theta}_0)(1 - \widehat{a}n^{-1}) \le z\big] = \mathbb{P}(\chi_q^2 \le z) + O(n^{-2}),$$

for any fixed $z > 0$. The estimator $\widehat{a}$ is specified in Section 2.4 of DiCiccio, Hall and Romano (1991). This is a remarkable advantage of EL: a simple correction for the mean of the empirical log-likelihood ratio reduces coverage error from order n^{-1} to order n^{-2}. It is known that the bootstrap is not Bartlett-correctable, and so the coverage accuracy of bootstrap methods cannot be enhanced by a simple correction. Usually, the bootstrap can only be corrected by resorting to a more computer-intensive methods such as bootstrap iteration, e.g. Hall (1986) and Beran (1987). In one of Peter's latest papers, Chang and Hall (2015) showed that a single double bootstrap resampling is enough for achieving bias reduction in estimating such defined $\boldsymbol{\theta}_0$, but not so for converge accuracy improvement of the confidence regions of $\boldsymbol{\theta}_0$. The Bartlett correction of EL has been examined in more general model settings; see, for example, Chen and Cui (2006, 2007).

2.3 Pseudo-likelihood theory for EL

Hall (1990) studied the confidence regions constructed by EL in a pseudo-likelihood framework for the same parameter $\boldsymbol{\theta}_0$ defined as that in Section 2.2. Hall (1990) showed that: (i) upon subject to a location parameter, EL draws second-order correct contours for those of a pseudo-likelihood; (ii) EL regions may be adjusted for location so as to render them second-order correct; (iii) location-adjusted EL regions are Bartlett-correctable, in the sense that a single empirical scale correction applied to location-adjusted EL reduces coverage error by an order of magnitude. However, the location adjustment alters the form of the Bartlett correction. Hall (1990) also pointed out the connection and difference between the EL and bootstrap likelihood (Hall, 1987) for constructing confidence regions.

Following the Wilks' theorem, the confidence region for $\boldsymbol{\theta}_0$ based on EL is given by $\mathcal{R}_E = \{\boldsymbol{\theta} \in \mathbb{R}^q : \ell_E(\boldsymbol{\theta}) \leq \chi^2_{q,1-\alpha}\}$ where $\chi^2_{q,1-\alpha}$ is the $(1-\alpha)$-quantile of χ^2_q distribution. Since $\bar{\mathbf{X}}_n$ is the maximum likelihood estimator of $\boldsymbol{\mu}_0$, then $\widehat{\boldsymbol{\theta}}_n = \mathbf{K}(\bar{\mathbf{X}}_n)$ states the maximum likelihood estimator of $\boldsymbol{\theta}_0$. Let $\widehat{\mathbf{Q}}_n$ be an estimator of the asymptotic variance matrix $\mathbf{Q}$ of $n^{1/2}\widehat{\boldsymbol{\theta}}_n$, and put $\widehat{\boldsymbol{\eta}}_n = \widehat{\mathbf{Q}}_n^{-1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$. If the density f of $\widehat{\boldsymbol{\eta}}_n$ were known, then the confidence region for $\boldsymbol{\theta}_0$ can be selected as the pseudo-likelihood region $\{\boldsymbol{\theta} \in \mathbb{R}^q : f\{\widehat{\mathbf{Q}}_n^{-1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\} \geq \nu\}$ for ν satisfying $\int_{\{\mathbf{y}\in\mathbb{R}^q : f(\mathbf{y})\geq v\}} f(\mathbf{y})d\mathbf{y} = 1-\alpha$. In practice, the form of f is usually unknown, so that assumption is often made that f is approximately equal to the multivariate standard normal density function ϕ. In this case, the pseudo-likelihood based confidence region becomes $\mathcal{R}_{P,1} = \{\boldsymbol{\theta} \in \mathbb{R}^q : |\widehat{\mathbf{Q}}_n^{-1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)|_2 \leq \delta\}$ for δ satisfying $\int_{\{\mathbf{y}\in\mathbb{R}^q : |\mathbf{y}|_2\leq\delta\}} \phi(\mathbf{y})d\mathbf{y} = 1-\alpha$. This approach fails to take into account the skewness, kurtosis, and other distributional information of $\widehat{\mathbf{Q}}_n^{-1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ and produces elliptical confidence regions. Therefore, not surprisingly, the contours of $\mathcal{R}_{P,1}$ and $\mathcal{R}_E$ do not agree with each other in the first-order. Hall (1990) proved that EL draws contours which are second-order correct for pseudo-likelihood contours based on $\widehat{\boldsymbol{\xi}}_n + n^{-1}\boldsymbol{\psi}$, instead of $\widehat{\boldsymbol{\eta}}_n$, where $\widehat{\boldsymbol{\xi}}_n = (\mathbf{Q}^{1/2}\widehat{\mathbf{Q}}_n^{-1}\mathbf{Q}^{1/2})^{1/2}\mathbf{Q}^{-1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ and $\boldsymbol{\psi}$ is a fixed vector-valued parameter. The exact form of $\boldsymbol{\psi}$ is given in Equation (3.15) of Hall (1990). Here "second-order correct" means that distance between contours at the same probability level drawn by the two different methods are $n^{-3/2}$ apart and they agree with each other up to the order n^{-1}. Since $\boldsymbol{\psi}$ is usually nonzero, the contours based on the $\widehat{\boldsymbol{\xi}}_n$ of the pseudo-likelihood are not second-order correct. Nevertheless, the contours are readily location-adjustable by re-centering. Let $\mathcal{R}_A = \mathcal{R}_E + n^{-1}\widehat{\mathbf{Q}}_n^{1/2}\widehat{\boldsymbol{\psi}}$ where $\widehat{\boldsymbol{\psi}}$ is an estimate of $\boldsymbol{\psi}$ satisfying $\widehat{\boldsymbol{\psi}} = \boldsymbol{\psi} + O_p(n^{-1/2})$. Denote by h the density of $n^{1/2}\widehat{\boldsymbol{\xi}}_n$ and let $\widehat{\mathbf{H}} = (\mathbf{Q}^{1/2}\widehat{\mathbf{Q}}_n^{-1}\mathbf{Q}^{1/2})^{1/2}\mathbf{Q}^{-1/2}$. The likelihood-based confidence region for $\boldsymbol{\theta}_0$, founded on the distribution of $\widehat{\boldsymbol{\xi}}_n$, is $\mathcal{R}_{P,2} = \{\boldsymbol{\theta} \in \mathbb{R}^q : -2\log[(2\pi)^{q/2}h\{n^{1/2}\widehat{\mathbf{H}}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\}] \leq \chi^2_{q,1-\alpha}\}$. Hall (1990) showed that the boundary of $\mathcal{R}_A$ is $O_p(n^{-3/2})$ away from that of $\mathcal{R}_{P,2}$. In addition, Hall (1990) proved that the location-adjusted confidence region $\mathcal{R}_A$ is also Bartlett-correctable in the sense that $\mathcal{R}_{AB} = \{\boldsymbol{\theta} + n^{-1/2}\widehat{\mathbf{Q}}_n^{1/2}\widehat{\boldsymbol{\psi}} : \ell_E(\boldsymbol{\theta}) \leq (1+\widehat{b}n^{-1})\chi^2_{q,1-\alpha}\}$ has coverage probability $1-\alpha+O(n^{-2})$ for some $\widehat{b}$.

2.4 Smoothed EL confidence intervals for quantiles

Peter's three papers reviewed in Section 2.1–2.3 all investigated the case where the statistic of interest is a smooth function of means. Chen and Hall (1993) studied the performance of EL in constructing confidence intervals for quantiles, which is a more sophisticated case. As pointed out by Owen (1988), when EL was used to construct confidence intervals for a population quantile, it reproduces precisely the so-called sign-test or binomial-method interval. A main disadvantage of the sign-test method is that the coverage error of the confidence intervals is usually $O(n^{-1/2})$ and it is hard to improve. The main contribution of Chen and Hall (1993) is that, by replacing the standard EL by a smoothed version, the coverage errors of the corresponding confidence intervals for quantiles reduce to $O(n^{-1})$, and it may be Bartlett-corrected to produce confidence intervals with coverage errors of only $O(n^{-2})$.

Let $X_1, \ldots, X_n$ denote a random sample from some distribution F. For a given $\beta \in (0, 1)$, we assume the βth quantile, $\theta_0 = F^{-1}(\beta)$, is uniquely defined, and we are interested in constructing its confidence intervals. Let $\mathcal{K}(\cdot)$ denote an mth-order kernel for some $m \geq 2$, and define $G(x) = \int_{y<x} K(y)dy$. Set $G_h(x) = G(x/h)$ for some bandwidth $h > 0$. For each $\theta \in (0, 1)$, define the EL for θ as $L_n(\theta) = \sup\{\prod_{i=1}^n w_i : w_i \geq 0,\ \sum_{i=1}^n w_i = 1,\ \sum_{i=1}^n w_i G_h(\theta - X_i) = \beta\}$. The associated empirical log-likelihood is $\ell_E(\theta) = -2\log\{L_n(\theta)\} - 2n\log n$. Chen and Hall (1993) showed that the Wilks' theorem still holds for $\ell_E(\boldsymbol{\theta}_0)$ in the sense $\ell_E(\theta_0) \to_d \chi_1^2$ as $n \to \infty$. The confidence interval for θ_0 is $\mathcal{I}_1 = \{\theta \in \mathbb{R} : \ell_E(\theta) \leq \chi^2_{1,1-\alpha}\}$. With some regularity conditions and suitable selection of the bandwidth h, Theorem 3.2 of Chen and Hall (1993) indicates that the coverage error of $\mathcal{I}_1$ is of order n^{-1}. As for the Bartlett correction, define $\mathcal{I}_2 = \{\theta \in \mathbb{R} : \ell_E(\theta) \leq (1 + \widehat{c}n^{-1})\chi^2_{1,1-\alpha}\}$ where $\widehat{c} = \widehat{\kappa}_2^{-2}\widehat{\kappa}_4/2 - \widehat{\kappa}_2^{-3}\widehat{\kappa}_3^2/3$ with $\widehat{\kappa}_j = n^{-1}\sum_{i=1}^n\{G_h(\widehat{\theta}_n - X_i) - \beta\}^j$ and $\widehat{\theta}_n$ is the usual estimate of θ_0. With suitable selection of the bandwidth h, the coverage error of $\mathcal{I}_2$ is of order n^{-2}; see Theorem 4.1 of Chen and Hall (1993) for details.

2.5 EL confidence bands in density estimation

Peter's last EL paper (Hall and Owen, 1993) was dedicated to construct the confidence bands in problems of nonparametric density estimation. As mentioned in Hall and Owen (1993), using the bootstrap to construct confidence bands from a large collection of simulated curve-estimates has several problems: (i) ambiguity due to that there are many "95% confidence bands" that contain, as envelopes, precisely 95% of the simulated estimates; (ii) to avoid ambiguity, the choice of point-wise confidence band with simultaneous 95% coverage probability may be practically problematic and/or computationally demanding from various aspects. and (iii) the resulting bands may not be smooth by connecting resampled curves. Hall and Owen (1993) illustrated how to apply EL in constructing the confidence bands for density functions so that the aforementioned problems are avoided.

Consider a sample $X_1, \ldots, X_n$ from some population with the density function f_0. The kernel density estimate of f_0 is given by $\widehat{f}(x) = (nh)^{-1}\sum_{i=1}^n \mathcal{K}_h(x - X_i)$, where $\mathcal{K}(\cdot)$ is a kernel function and $\mathcal{K}_h(x) = \mathcal{K}(x/h)$. Suppose that we are interested in bands over the interval $[0, 1]$. Let $K_i(x) = h^{-1}\mathcal{K}_h(x - X_i)$ for $i = 1, \ldots, n$ and $x \in [0, 1]$. Define $\mu_0(x) = \mathbb{E}\{K_i(x)\}$ for any $x \in [0, 1]$. For a function f_1 defined on $[0, 1]$ satisfying $\min_{1\leq i\leq n} K_i(x) \leq f_1(x) \leq \max_{1\leq i\leq n} K_i(x)$ for any $x \in [0, 1]$, the empirical log-likelihood is given by $\ell_E(f_1) = \ell_E(f_1)(x) = -2\log\{L_n(f_1)/L_n(\widehat{f})\} = 2\sum_{i=1}^n \log[1 + \lambda(x)\{K_i(x) - f_1(x)\}]$, where $\lambda(x)$ is the Lagrange multiplier satisfying $\sum_{i=1}^n\{K_i(x) - f_1(x)\}[1 + \lambda(x)\{K_i(x) - f_1(x)\}]^{-1} = 0$. Similarly, define $s(f_1)(x) = \mathrm{sgn}\{\widehat{f}(x) - f_1(x)\}\{\ell_E(f_1)(x)\}^{1/2}$. Given the numbers $c, c_1, c_2 > 0$, define the classes $F_1(c) = \{f :$

$\ell(f)(x) \leq c$ for any $x \in [0,1]\}$ and $F_2(c_1, c_2) = \{f : c_1 \leq s(f)(x) \leq c_2$ for any $x \in [0,1]\}$. Hall and Owen (1993) took these classes, with appropriate values of c, or c_1 and c_2, as confidence set for μ_0 over $[0,1]$. Hall and Owen (1993) suggested to use bootstrap calibration to determine c, c_1 and c_2. More specifically, one draws a resample $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ from the sample $\mathcal{X} = \{X_1, \ldots, X_n\}$, using random sampling with replacement. Then the bootstrap empirical log likelihood is constructed as $\ell_E^*(f_1) = \ell_E^*(f_1)(x) = 2\sum_{i=1}^n \log[1 + \lambda^*(x)\{K_i^*(x) - f_1(x)\}]$, with $K_i^*(x) = h^{-1}\mathcal{K}_h(x - X_i^*)$ and $\lambda^*(x)$ satisfying $\sum_{i=1}^n \{K_i^*(x) - f_1(x)\}[1 + \lambda^*(x)\{K_i^*(x) - f_1(x)\}]^{-1} = 0$. Analogously, $s^*(f_1) = \mathrm{sgn}\{\widehat{f}^* - f_1\}\{\ell_E^*(f_1)\}^{1/2}$ where $\widehat{f}^* = (nh)^{-1}\sum_{i=1}^n K_i^*$. Then one can calculate $\widehat{c}$, $\widehat{c}_1$ and $\widehat{c}_2$ by ensuring $\mathbb{P}\{\ell_E^*(\widehat{f})(x) \leq \widehat{c}$ for any $x \in [0,1] \mid \mathcal{X}\} = 1 - \alpha$ and $\mathbb{P}\{\widehat{c}_1 \leq s^*(\widehat{f})(x) \leq \widehat{c}_2$ for any $x \in [0,1] \mid \mathcal{X}\} = 1 - \alpha$. Since $f_0 = \mu_0 + \text{bias}$, with the standard approach in nonparametric density estimation, one can obtain the estimate, denoted by $\widehat{\beta}$, for the bias term $f_0 - \mu_0$. Then $F_1(\widehat{c}) + \widehat{\beta}$ and $F_2(\widehat{c}_1, \widehat{c}_2) + \widehat{\beta}$ provide two confidence bands for f_0.

3 Current challenges for EL

We have reviewed Peter's contribution to EL. In summary, Peter has done fundamental work on investigating the coverage accuracy of confidence regions constructed with EL. Peter's work deepens the understanding of the properties of EL, and broadens the scope of EL as an important nonparametric device for statistical inference. In the current paradigm of problems with increasing dimensionality and complexity, EL as a computer-intensive nonparametric device is facing two main challenges from two questions – how are the properties of EL for statistical inferences with those new problems, and how can the computational cost be handled? The same challenges are also posed for other nonparametric devices like the bootstrap.

Coverage accuracy, both theoretically and practically, is crucial problem. Increasing data dimensionality is indeed a very challenging problem for EL. As pointed out in Tsao (2004), when the data dimensionality is moderately large but fixed, the empirical log-likelihood ratio has a non-zero probability to be infinity, so that under-coverage occurs especially when the sample size is small. Tsao and Wu (2013, 2014) proposed extended EL to improve the coverage accuracy, which are first order equivalent to the conventional EL, and is also second order accurate with the Bartlett correction. Related to the under-coverage and the so-called empty set problem (no solution exists for the constrained optimization in EL), Chen, Variyath and Abraham (2008) and Emerson and Owen (2009) proposed adjustments with adding extra data points so that a valid solution always exists.

It is more challenging when considering EL with high-dimensional data in the sense that the dimensionality of the parameter is allowed to diverge with the sample size. Hjort, McKeague and Van Keilegom (2009) and Chen, Peng and Qin (2009) studied the properties of EL, allowing the number of parameters to diverge at some polynomial rate with the sample size. Their results show that asymptotically, the limiting distribution of the empirical log-likelihood ratio can still be characterized by the χ^2 distribution in the sense that with appropriate scaling and normalization, the empirical log-likelihood ratio converges in distribution to the standard normal distribution when the number of parameters is diverging.

Tang and Leng (2010) and Leng and Tang (2012) introduced penalty functions on the magnitudes of the parameters with the purpose to produce sparse model estimations. Their frameworks also consider that the number of parameters is allowed to diverge with the sample size at some polynomial rate of the sample size. Lahiri and Mukhopadhyay (2012) considered a different formulation EL with penalization that can accommodate higher dimensional model parameters that can exceed the sample size. Their definition of EL is different and the penalty is introduced as a deviation function of the parameter defined. Chang, Chen and Chen (2015) considered the impact of dependence on EL and penalized EL with diverging number of model parameters, and demonstrated the validity and properties of EL in an extended framework including time series and other dependent data.

Chang, Tang and Wu (2013, 2016) considered EL as a tool for the sure variable screening purpose in the sense of Fan and Lv (2008) when dealing with high-dimensional data. By using the idea of treating variables marginally using EL, Chang, Tang and Wu (2013, 2016) showed that the empirical log-likelihood ratio evaluated at zeros are very informative for detecting contributing variables in linear models, generalized linear models, and a class of nonparametric and semiparametric models including the single-index models and varying-coefficient models.

Most recently, Shi (2016) studied a formulation of EL with additional constraint from the magnitudes of the conditions in EL. The target problem of Shi (2016) is inferences for a fixed dimensional model parameter with many moment conditions that may grow exponentially with sample size. Results in Shi (2016) show that EL can accommodate high-dimensional moment conditions by relaxing the equality constraints to some inequality constraints. Targeting at estimating a sparse high-dimensional model parameter whose size may grow exponentially with the sample size, Chang, Tang and Wu (2017a) considered a new formulation of penalized EL regularizing both the magnitudes of the parameter and the Lagrange multiplier induced by EL. Chang, Tang and Wu (2017b) studied the statistical inferences problems including confidence set estimation and specification test in such setting.

Though some progress has been made for EL to meet the current challenges, there are more problems remain open. For example, there is not yet a counterpart in high-dimensional cases of the most attractive merit of EL being self studentized so that the empirical log-likelihood ratio can be conveniently applied for statistical inferences with good coverage properties as established by Peter' work in conventional cases. Additionally, practically it is also very important to explore on how to more efficiently conduct the optimizations with high-dimensional model parameter and moment conditions.

4 Personal reminisces

We first met Peter when we were PhD students in Statistics. As an experience shared by many others, and maybe even as a tradition, we were immediately left with the impression that Peter is such a friendly and easily going gentlemen, being so supportive, helpful, and exemplary to younger generations like us. In those years after, we saw that familiar back of the figure with his head down and glasses off working with his laptop on a table in the middle of a crowded conference hall, no difference from when he was in a quite office in a university campus. There we understand what is beyond the productivity of those prominent works, and also why our email requests were almost immediately attended no matter where he was, no matter what exactly the time was.

We had greatly benefited from Peter's wisdom. Peter's contribution to EL has direct influence in our works in the area. Peter's investigations and analyses always shed lights on the idea, direction, and path for solving our problems. Peter's advice and influence, from our direct contact with him, and from the indirect transferred from our adviser and colleagues, now becomes our most precious fortune. Peter, we miss you!

Jinyuan Chang would like to take this opportunity to appreciate the supports from Peter during his stay at the University of Melbourne. As the last postdoc of Peter, Peter gave him a lot of freedom for conducting independent research of his interest. Jinyuan was invited to give an on-site interview at London School of Economics for his faculty position application in the Statistics department in early February 2015. Due to family issues, Jinyuan thought that maybe he cannot take this position and preferred to quit the interview. He wrote an email to Peter for this in January 10 2015 (Saturday). Peter responded immediately and told him "don't burn any bridges". Peter even came to office on Sunday (January 11 2015) and had a chat with Jinyuan in person. At that moment Peter was struggling against the illness and had not gone to office for a while. Jinyuan was deeply touched by the kindness and consideration from Peter. Indeed, Peter was always a kind and warm-hearted person who will be dearly missed and forever remembered.

Acknowledgments

Jinyuan Chang is supported in part by the Fundamental Research Funds for the Central Universities (Grant No. JBK170161, JBK150501), NSFC (Grant No. 11501462), the Center of Statistical Research at SWUFE, and the Joint Lab of Data Science and Business Intelligence at SWUFE. Jinyuan Chang would like to also thank the supports and helps from Prof. Aurore Delaigle after Peter passed away. Cheng Yong Tang is supported in part by the NSF Grants IIS-1546087 and SES-1533956.

References

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika*, **74**, 457–468.

Chang, J., Chen, S. X. and Chen, X. (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, **185**, 283–304.

Chang, J. and Hall, P. (2015). Double-bootstrap methods that use a single double-bootstrap simulation. *Biometrika*, **102**, 203–214.

Chang, J., Tang, C. Y. and Wu, T. T. (2017a). A new scope of penalized empirical likelihood with high-dimensional estimating equations. Available at arXiv:1704.00566.

Chang, J., Tang, C. Y. and Wu, T. T. (2017b). High-dimensional statistical inferences with over-identification: confidence set estimation and specification test. *Manuscript*.

Chang, J., Tang, C. Y. and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *The Annals of Statistics*, **41**, 2123–2148.

Chang, J., Tang, C. Y. and Wu, Y. (2016). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *The Annals of Statistics*, **44**, 515–539.

Chen, J., Mulayath, A. M. and Abraham, B. (2008). Adjusted empirical likelihood and its properties. *Journal of Computational and Graphical Statistics*, **17**, 426–443.

Chen, S. X. and Cui, H. (2006). On Bartlett correction of empirical likelihood in the presence of nuisance parameters. *Biometrika*, **93**, 215–220.

Chen, S. X. and Cui, H. (2007). On the second properties of empirical likelihood with moment restrictions. *Journal of Econometrics*, **141**, 492–516.

Chen, S. X. and Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, **21**, 1166–1181.

Chen, S. X., Peng, L. and Qin, Y. L. (2009). Effects of data dimension on empirical likelihood. *Biometrika*, **96**, 711–722.

Chen, S. X. and Van Keilegom, I. (2009). A review on empirical likelihood methods for regression. *Test*, **18**, 415–447.

DiCiccio, T., Hall, P. and Romano, J. (1989). Comparison of parametric and empirical likelihood functions. *Biometrika*, **76**, 465–476.

DiCiccio, T., Hall, P. and Romano, J. (1991). Empirical likelihood is bartlett-correctable. *The Annals of Statistics*, **19**, 1053–1061.

Emerson, S. C. and Owen, A. B. (2009). Calibration of the empirical likelihood method for a vector mean. *Electronic Journal of Statistics*, **3**, 1161–1192.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society*, **B**, **70**, 849–911.

Hall, P. (1986). On the bootstrap and confidence intervals. *The Annals of Statistics*, **14**, 1431–1452.

Hall, P. (1987). On the bootstrap and likelihood-based confidence regions. *Biometrika*, **74**, 481–493.

Hall, P. (1990). Pseudo-likelihood theory for empirical likelihood. *The Annals of Statistics*, **18**, 121–140.

Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *International Statistical Review*, **58**, 109–127.

Hall, P. and Owen, A. B. (1993). Empirical likelihood confidence bands in density estimation. *Journal of Computational and Graphical Statistics*, **2**, 273–289.

Hjort, N. L., McKeague, I. and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *The Annals of Statistics*, **37**, 1079–1111.

Lahiri, S. N. and Mukhopadhyay, S. (2012). A penalized empirical likelihood method in high dimensions. *The Annals of Statistics*, **40**, 2511–2540.

Leng, C. and Tang, C. Y. (2012). Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika*, **99**, 703–716.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, **18**, 90–120.

Owen, A. B. (2001). *Empirical Likelihood.* Chapman and Hall-CRC, New York.

Shi, Z. (2016). Econometric estimation with high-dimensional moment equalities. *Journal of Econometrics*, **195**, 104–119.

Tang, C. Y. and Leng, C. (2010). Penalized high dimensional empirical likelihood. *Biometrika*, **97**, 905–920.

Thomas, D. R. and Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, **70**, 865–871.

Tsao, M. (2004). Bounds on coverage probabilities of the empirical likelihood ratio confidence regions. *The Annals of Statistics*, **32**, 1215–1221.

Tsao, M. and Wu F. (2013). Empirical likelihood on the full parameter space. *The Annals of Statistics*, **41**, 2176–2196.

Tsao, M. and Wu F. (2014). Extended empirical likelihood for estimating equations. *Biometrika*, **101**, 703–710.

Zhou, M. (2016). *Empirical Likelihood Method in Survival Analysis*. Chapman and Hall-CRC, New York.