A brief survey of numerical procedures for empirical likelihood

Rodolphe Priam

▶ **To cite this version:**

Rodolphe Priam. A brief survey of numerical procedures for empirical likelihood. 2021. hal-03095014

HAL Id: hal-03095014
https://hal.science/hal-03095014

Preprint submitted on 4 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A brief survey of numerical procedures for empirical likelihood*

R. Priam†

January 4, 2021

Abstract

Empirical likelihood (EL) is a statistical framework for non parametric parameters inference. It adds weights in an estimating equation in order to proceed jointly to the parameters inference and to the estimation of the weights via an additional criterion. This results into a new objective function depending on the parameters and the weights with embedded lagrangian multipliers. Thus algorithms have been proposed in the current literature in order to maximize this function under constraints. EL is very appealing because it is more general than the usual parametric likelihoods: EL is enough flexible to handle many kind of estimating equations. It inference appears more restrictive in comparison to parametric likelihood thus several approaches have been developed in order to leverage this limitation. This paper presents an overview of the existing inferential and numerical methods with computational perspectives.

1 Introduction

Empirical likelihood [1] is defined through the hypothesis of the existence of a finite-dimensional parameter of interest. This parameter comes from the empirical distribution function defined after an observed sample data. EL constructs a profile likelihood by maximizing a joint probability of the data, under constraints. This is a different approach than the other non-parametric models but with a direct application to survey theory: the parameter is typically a quantity related to the population [2, 3]. Asymptotic theory for empirical likelihood shares some similarities[4] with the more usual parametric likelihood [5, 6], because a likelihood-ratio test may be also relevant.

In parameter inference, when a sample is available, the model may be related to an estimating equation. For instance, in maximum likelihood, after having evaluated the derivatives of the log-likelihood may be called the normal equations but also the estimating equations. Sometimes it is possible to solve such equations directly to find the estimates of the true parameters, but this is not possible in general. A usual approch is a gradient ascent procedure for maximum likelihood or a Newton procedure for solving systems of nonlinear equations in the more general case. If these approaches are possible, a statistical approach, which introduces a weighting, is recently studied extensively because it is often able to provide more efficient estimators of the parameters. In empirical likelihood (EL) theory [1], one can estimate an unknown parameter vector θ by maximizing the empirical likelihood defined as,

$$L(\mathbf{w}) = \prod_{i=1}^{n} w_i \,,$$

subject to constraints involving either the vector θ and either no parameters, under the form $\sum_i w_i \phi_i$ where ϕ_i is a (multi-dimensional) function of θ or a constant. Generally, the weights are positive and sum to one. With $\mathbf{w} = (w_1, \cdots, w_n)'$, they belong to the set,

$$\mathcal{W}_\pi = \{\mathbf{w} \in [0,1]^n, \sum_{i=1}^{n} w_i = 1\}.$$

As an illustration, we present two examples of estimating equations, the mean estimation and the regression estimation which can be solved from [7] in a general case. In generalized linear models the parameters are found by solving the score equations: the normal equations obtained after differentiating the likelihood w.r.t. the parameters. A consequence is that the number of univariate equations to be solved simultaneously equals the number of parameters.

Example 1: *In mean estimation, the central parameter is found by solving for the unknown parameter θ which is an expectation such that when θ and x_i is a p-dimensional*

*A nearly similar version of this document was sent to review previously since 09/2017 at first to one computer journal and after to one statistical journal.

†rpriam@gmail.com

vectors[8],

$$\phi(x_i, \theta) = x_i - \theta\,.$$

This comes from the usual sample mean estimator which is retrieved by summing the n function $\phi(.,.)$ while equating to zero. □

Example 2: *Let have $Z_i = (Y_i, X_i)$ with associated sample units (y_i, x_i) where $x_i = (x_i^{(1)}, \ldots, x_i^{(k)}, \ldots, x_i^{(p)})'$, and the hidden parameter $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$. For the regular regression model[8, 9]:*

$$\phi(Z_i, \theta) = x_i\big\{y_i - (x_i'\beta)\big\}\,.$$

This comes from the derivative of the log-likelihood when the variance parameter is supposed known or just not a parameter of interest in a non parametric setting. □

In the current literature, the pioneering researches in the domain are extended to particular models where the former theory -such as in [7]- does not apply directly. Moreover, EL is also improved by decreasing the numerical pitfalls or issues coming from the constrained optimization. The optimization for the weights generally leads to a closed-form expression as a function of the parameters from the constraints and the vector of lagrangian multipliers. Herein we focus on an overview of the functional expressions of the weights w_i and also the algorithms that are generally implemented for their estimation. Existing surveys and reviews about empirical likelihood focus mainly on the theory while herein the contents focuses on the computational components. Herein the population is presumed to be infinite unless if its size is specified such as in the context of survey theory.

Currently, some theoritical papers such that [10] may contain a (small) paragraph about the inferential algorithms or their implementations[1] for empirical likelihood. But only a small number of papers in the current literature [11, 12, 13, 14, 15, 16, 17, 18] are really dedicated to these questions. This motivates the proposal of a short and unified introduction to computational and numerical approaches to derive the inference of EL and several of its generalizations as a new complement to older and/or theoritical existing reviews such as [19, 10, 11, 20, 21, 22]. The plan is organized as follows. Section 1 contains the introduction. Section 2 contains the main numerical algorithms from the literature on empirical likelihood, including in a second sub-section the choices for the optimization of the objective functions explained in a first subsection. Section 3 contains several generalizations of empirical likelihood with alternative objective functions for improving the results in comparison to the former maximization of a constrained sum of the logarithm of unknown weigths. Section 4 proposes several computational applications from the domain of parametric likelihood. Section 5 contains the conclusion and discussion.

[1] Typically in R, Python or Matlab languages.

2 Existing computational algorithms

Diverse algorithms have been proposed in the current literature for maximizing the empirical likelihood under constraints. For each $\theta \in \Theta$ where Θ is the set of possible parameters it is denoted:

$$\begin{array}{rcl} \mathcal{W}_\theta & = & \{\mathbf{w}; \sum_{i=1}^n w_i \phi(z_i, \theta) = 0_\ell\}\,, \\ \mathcal{W}_\gamma & = & \{\mathbf{w}; \sum_{i=1}^n w_i g(z_i) = \gamma\}\,. \end{array}$$

Here, the function $\phi(.,.)$ is supposed ℓ-dimensional with component $\phi_k(.,.)$ while the function $g(.)$ and the vector $\gamma = (\gamma_1, \cdots, \gamma_m)'$ are both m-dimensional. We also supposed that the set $\mathcal{W} = \bigcup_{\theta\in\Theta}(\mathcal{W}_\theta \bigcap \mathcal{W}_\gamma)$ is non-empty. Otherwise there does not exist a solution to the search for an optimal weighting.

Three cases are considered in this section, the nested optimization where ξ and θ are infered separaterly and with or not estimating equations not involving θ, and the two-steps optimization where the weights are first infered with the estimating equation not involving θ followed by the inference of θ with the weights previously found.

2.1 Nested maximization algorithm

In this subsection, we review EL without additional constraints first, followed by EL with aditional ones.

EL with only constraints for θ In the following, we consider a sample $\mathcal{D} = \{z_1, z_2, \cdots, z_n\}$ and an unknown parameter vector denoted θ_0. The method EL aims at estimate this quantity and deduce the confidence intervals of the estimators. We suppose that each p_i is the probability mass placed on X_i. Then the EL for paramater θ is defined as maximizing the function $L(w)$ subject to the restrictions for weights:

$$\mathbf{w} \in \mathcal{W}_\pi \text{ and } \mathbf{w} \in \mathcal{W}_\theta\,.$$

For θ fixed, a unique solution may exist, when 0 is inside the convex hull of the points $\phi(z_1, \theta)$, $\phi(z_2, \theta)$, $\cdots$, and $\phi(z_n, \theta)$. This maximum can be found by using suitable Lagrangian multipliers as follows.

Sketch of the proof (see [7]) *It is denoted a vector with ℓ components $\xi(\theta) = (\xi^{(1)}, \xi^{(2)}, \cdots, \xi^{(\ell)})$ for the functional constraints, and $\xi_{\mathbf{w}}$ for the summation to the unity of the quantities w_i. Then, it is defined for maximizing* $\log(L)$,

$$\begin{aligned} M(\theta, \xi_{\mathbf{w}}, \xi) &= \log L(\mathbf{w}) + \xi_{\mathbf{w}}(1 - \textstyle\sum_i w_i) \\ &\quad -n\xi' \textstyle\sum_i w_i \, \phi(z_i, \theta) \, . \end{aligned}$$

By computing the partial derivative of $M(.)$ w.r.t. w_i and and summing the n derivatives after multiplying each one by the weights, we get the following solution depending to the unknown value of θ, as:

$$w_i = w(z_i, \xi, \theta) = \frac{1}{n} \frac{1}{1 + \xi' \phi(z_i, \theta)} \, .$$

These multipliers are then such as:

$$R(\xi, \theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\phi(z_i, \theta)}{1 + \xi' \phi(z_i, \theta)} = 0_\ell \, ,$$

which permits to find Lagrange parameter ξ by nonlinear optimization, with θ given. □

It has been shown in the current literature that the function $\theta \to \xi(\theta)$ is a continuous fonction of its parameters by the inverse function theorem. The EL problem is also equivalent to the minimization of the empirical log-likelihood ratio:

$$r(\xi, \theta) = \sum_{i=1}^{n} \log \left\{ w(z_i, \xi, \theta) \right\} \, .$$

This ratio is optimized to get the EL estimator denoted $\tilde{\theta}$ which estimates the parameter θ, and called maximum empirical likelihood estimate. Note that,

$$R(\xi, \theta) = \partial r(\xi, \theta) / \partial \xi \, .$$

Similarly the derivative with respect of the parameters θ leads to an other normal equation. With the additional notation at the introduction of the section, for this first approach, the empirical likelihood estimation leads to a nested maximization algorithm by solving for:

$$\hat{\theta} \;=\; \text{argmax}_{\theta \in \Theta} \left\{ \max_{\mathbf{w} \in \mathcal{W}_\pi \bigcap \mathcal{W}_\theta} L(\mathbf{w}) \right\} \, .$$

Some approaches exist in the current literature when there are additional constraints as presented hereafter.

EL with additional constraints When the constraints on the weights coming from the estimating equation are completed with additional constraints not depending on the parameter θ, the estimate of θ belongs to a more restrictive set of possible values:

$$\mathbf{w} \in \mathcal{W}_\pi \text{ and } \mathbf{w} \in \mathcal{W}_\theta \text{ and } \mathbf{w} \in \mathcal{W}_\gamma \, .$$

When we denote $\phi(.) = g(.) - \gamma$, a similar approach than above leads to the new expression after solving for the new multipliers ξ and ξ_γ as follows,

$$w_i = w(z_i, \xi, \xi_\gamma, \theta) = \frac{1}{n} \frac{1}{1 + \xi' \phi(z_i, \theta) + \xi_\gamma' \phi(z_i)} \, .$$

The new likelihood ratio is then,

$$r(\xi, \xi_\gamma, \theta) = \textstyle\sum_{i=1}^{n} \log \left\{ w(z_i, \xi, \xi_\gamma, \theta) \right\} \, .$$

The multipliers are then such as:

$$\begin{array}{l} \frac{1}{n} \sum_{i=1}^{n} \frac{\phi(z_i, \theta)}{1 + \xi' \phi(z_i, \theta) + \xi_\gamma' \phi(z_i)} = 0_\ell \\ \frac{1}{n} \sum_{i=1}^{n} \frac{\phi(z_i)}{1 + \xi' \phi(z_i, \theta) + \xi_\gamma' \phi(z_i)} = 0_m \, . \end{array}$$

In a matricial notation this can be seen in a more compact view as. The multipliers are then such as:

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1}{1 + \begin{bmatrix} \xi \\ \xi_\gamma \end{bmatrix}' \begin{bmatrix} \phi(z_i, \theta) \\ \phi(z_i) \end{bmatrix}} \begin{bmatrix} \phi(z_i, \theta) \\ \phi(z_i) \end{bmatrix} \right\} = \begin{bmatrix} 0_\ell \\ 0_m \end{bmatrix} \, .$$

With the additional notation at the introduction of the section, for this first approach, the empirical likelihood estimation leads to a nested maximization algorithm by solving for:

$$\hat{\theta} \;=\; \text{argmax}_{\theta \in \Theta} \left\{ \max_{\mathbf{w} \in \mathcal{W}_\pi \bigcap \mathcal{W}_\theta \bigcap \mathcal{W}_\gamma} L(\mathbf{w}) \right\} \, .$$

Some approaches exist in the current literature, with a less direct maximization. Note that usually authors in the current literature consider a modified Newton-Raphson procedure (see [2, 11, 12]) for solving the optimization problems above.

2.2 Two-steps algorithm for additional constraints

In [23] (see also [24]) the authors consider a sample of individuals which can be completed with information at the

population level for modeling a dependent variable and explanatory variables. This information can reduce bias and decrease the variance of the parameter estimates. Formally, the authors add this information via linear constraints on the functions of the model in an empirical likelihood framework for generalized linear models. This approach is inspired from data combination in econometrics. They propose a two-step which generalizes the method-of-moments from [25]. The authors propose to solve for θ and $\mathbf{w}$ via the empirical profile likelihood estimator of [7] or via a two steps procedure. The first approach is described in the subsection above.

The second approach, which is the two-steps estimation procedure, leads to a simpler computational algorithm and easier theoretical study. This is defined as follows. First, the weights are found with θ given, then θ is found with the weigths given. The steps are iterated until convergence to a stable value. This is written as follows, where $n\,\mathbf{w}$ is the vector of components $n\,w_i$,

$$\begin{array}{rcl} \hat{\mathbf{w}} & = & \text{argmax}_{\mathbf{w}\in\mathcal{W}_\pi \bigcap \mathcal{W}_\gamma} \log\left(L(n\,\mathbf{w})\right) \\ \hat{\theta} & \in & \{\theta\in\Theta, \sum_{i=1}^n \hat{w}_i \phi(z_i,\theta) = 0_\ell\} \end{array}$$

The first step can be solved via the standard algorithm in EL but the weights do not depend on θ anymore contrary to the procedure described in the previous section. The second step asks for an algorithm dedicated to solve an (generalized) estimating equation. If the two steps has no acceptable solution, the previous nested maximization presented in the subsection above is required. For the two steps procedure, they propose the explicit expressions in the case of the generalized linear models. The constraints are supposed m-dimensional with $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_m)'$ and $g = (g_1, g_2, \ldots, g_m)'$, such that is defined $\phi_j(.) = g_j(.) - \gamma_j$. The weights are estimated as,

$$\hat{w}_i = w(z_i, \hat{\xi}_\gamma) = \frac{1}{n}\frac{1}{1+\sum_{j=1}^m \hat{\xi}_\gamma^{(j)}\phi_j(z_i)}\,,$$

with the estimated $\xi_\gamma^{(j)}$ equal to $\hat{\xi}_\gamma^{(j)}$, and by writting the weigths estimates as a function of z_i, γ, ξ_γ. Finally, with a vectorial notation this leads for θ to the following normal equation:

$$\tfrac{1}{n}\sum_{i=1}^n \frac{1}{1+\hat{\xi}_\gamma'\phi(z_i)}\phi(z_i,\theta) = 0_\ell\,.$$

Note that for the two-steps approach, the algorithm and the variances formula have been implemented in an R package for the generalized linear models.

Algorithm 1: NR procedure for EL [16, 1] with log(.)

Result: The estimate for the vector ξ
Input : ξ_0, $t=0$, $\gamma_0=1$, $\epsilon = 10^{-8}$
Output: $\hat{\xi}$

$\Delta_1(\xi^{(0)}) = \sum_i \frac{1}{1+\phi_i'\xi^{(0)}}\,\phi_i$
$\Delta_2(\xi^{(0)}) = -\sum_i \frac{1}{(1+\phi_i'\xi^{(0)})^2}\,\phi_i\phi_i'$
repeat
  $\delta^{(t)} = \gamma^{(t)}\Delta_2(\xi^{(t)})$
  while $\exists i,\ 1+\xi^{(t)} - \phi_i'\delta^{(t)} \le 0$ **do**
    $\gamma^{(t)} = \gamma^{(t)}/2$
    $\delta^{(t)} = \gamma^{(t)}\Delta_2(^{(t)})$
  end
  $\xi^{(t+1)} = \xi^{(t)} - \delta^{(t)}$
  $\gamma^{(t+1)} = \frac{1}{\sqrt{t+1}}$
  $t = t+1$
  $\Delta_1(\xi^{(t)}) = \sum_i \frac{1}{1+\phi_i'\xi^{(t)}}\,\phi_i$
  $\Delta_2(\xi^{(t)}) = -\sum_i \frac{1}{(1+\phi_i'\xi^{(t)})^2}\,\phi_i\phi_i'$
until $|\Delta_2(\xi_k)| \le \epsilon$
$\hat{\xi} = \xi^{(t)}$.

2.3 Numerical algorithms

For numerical experiments, empirical likelihood appears challenging because of the nonlinearity and the constraints. The objective function is convex for the ξ but may be highly nonlinear for the parameter θ with nonconvexity and eventual multiple local solutions. For small number of variables, some authors propose the Nelder-Mead algorithm [26] called simplex method. Note that for non smooth functions, some approaches extends the theory of empirical likelihood for allowing the estimation[27].

NR algorithms: Pioneer algorithms [2, 11, 12, 15, 16] for EL are based on the Newton-Raphson (NR) procedure with the projected version for the constraints. In particular, the denominator in the quotient for w_i is required to remain positive and non null as a supplementaty condition to the linear constraints generally introduced, as found in Algorithm 1. Hence when t stands for the time step, this algorithm implements the iterative procedure where the derivatives are evaluated given the last current value of the parameter θ of

Algorithm 2: NCD procedure for EL [16] with $\log_*(.)$

Result: The estimate for the vector ξ
Input : ξ_0, $t=0$, $\gamma_0=1$, $\epsilon=10^{-8}$
Output: $(\hat{\xi},\hat{\theta})$

$t=0$
repeat
  for $1 \leq j \leq d$ **do**
    $\theta_j^{(t+1)} = \theta_j^{(t)} - \frac{A_j}{B_j+C_j}$
    $A_j = \sum_i \log_*'(s_i^{(t)})[\xi' \frac{\partial \phi_i(\theta)}{\partial \theta_j}]$
    $B_j = \sum_i \log_*''(s_i^{(t)})[\xi' \frac{\partial \phi_i(\theta)}{\partial \theta_j}]^2$
    $C_j = \sum_i \log_*'(s_i^{(t)})[\xi' \frac{\partial^2 \phi_i(\theta)}{\partial \theta_j^2}]$
    for $1 \leq k \leq m$ **do**
      $\xi_k^{(t+1)} = \xi_k^{(t)} - \frac{\sum_i \log_*'(u_i^{(t)})[\frac{\partial \phi_{ik}(\theta)}{\partial \theta_j}]}{\sum_i \log_*''(u_i^{(t)})[\frac{\partial \phi_{ik}(\theta)}{\partial \theta_j}]^2}$
    end
  end
  $t=t+1$
until $max_{1\leq j\leq d}|\theta_j^{(t)} - \theta_j^{(t-1)}| \leq \epsilon$
$\hat{\xi}=\xi^{(t)}$ and $\hat{\theta}=\theta^{(t)}$.

interest:

$$\xi^{(t+1)} = \xi^{(t)} - \gamma_t \left[\frac{\partial^2 \log L(\mathbf{w})}{\partial \xi \partial \xi'}\bigg|_{\xi=\xi^{(t)}} \right]^{-1} \frac{\partial \log L(\mathbf{w})}{\partial \xi}\bigg|_{\xi=\xi^{(t)}}.$$

Note that the logarithm function $\log(.)$ is typically replaced by an alternative continuous alternative named[2] $\log_*(.)$, with better numerical behaviors and able to keep the asymptotics properties of empirical likelihood. When the quadratic approximation leads to bad fitting, it has been proposed an alternative function with a fourth-order polynomial version instead of the previous second order one in [28]. When the function $z \to \log_*(z)$ is considered, numerically more robust alternative expressions for the first order derivative $\Delta_1(\xi_k)$ and the second order derivative $\Delta_2(\xi_k)$ have their formula given in [16] and in chapter 12 of [1]. It is also proposed to implement a MM algorithm by replacing the hessian with a more constant matrix. Note that in the case[3] of a Newton algorithm the optimization for θ may be very similar if the two steps algorithms is not considered.

Moreover, a more numerical stable algorithm based on the nested coordinate descent procedure is proposed recently in [16] by component-wise update, see also [29]. The algorithm was stated as named here Algorithm 2, with $s_i^{(t)} = 1 + \phi_i' \xi^{(t)}$ and $u_i^{(t)} = 1 + \xi' \phi_i^{(t)}$ where $\phi_i^{(t)}$ is evaluated at $\theta^{(t)}$. It implements a element-wise Newton-Raphson algorithm, with derivatives also evaluated at the last value of the component of the parameter of interest. For all components k and j, this is written as:

$$\theta_j^{(t+1)} = \theta_j^{(t)} - \frac{\partial \log L(\mathbf{w})/\partial \xi_j}{\partial^2 L(\mathbf{w})/\partial \xi_j^2}.$$

$$\xi_k^{(t+1)} = \xi_k^{(t)} - \frac{\partial \log L(\mathbf{w})/\partial \xi_k}{\partial^2 L(\mathbf{w})/\partial \xi_k^2}.$$

Same kind of keep step size is introduced in practice as in the Newton-Raphson procedure to insure the constraints of positivity of the probabilities $\hat{w}_i$ in Algorithm 1. Note that it is in general presented the joint update of the quantities ξ and θ as not stable ine the case of nonlinear functions ϕ_i and to be avoided, otherwise a similar stategy may be considered in this risky approach.

Perspectives: As convex optimization is currently extensively studied, the pioneering procedures presented in Algorithm 1 and 2 may be adapted, in particular to speed up their running time eventually for the larger samples and the larger number of variables. For convexity and optimization, interior point methods with logarithmic barriers [30] or trust region Newton method [31] are worth to tried but may be more complex to implement in practice. Moreover, the case when a penalty is added for ξ is exactly intensively studied as in [32, 33] or also DC programming [34] for instance. Several authors have prefered to proceed with the help of a relevant numerical solver without directly implementing the algorithm for the optimization. Some existing packages sometimes proprietary are also to solve this problem for academic purpose. Even for the weights, a solution is available directly via software as for calibration without empirical likelihood in [35] or direct mutiplicative updates [36] for instance. In conclusion, for direct or open source implementations, as pointed in [15], there exists alternative procedures to the usual Newton-Raphson one for convex optimization, hence they may be used more often in future in order to improve the final estimates even further.

[2] This function $z \to \log_*(z)$ equals the logarithmic one for $z \geq n^{-1}$ and a quadratic one otherwise equal to $\log(n^{-1}) + 2nz - 0.5n^2z^2$.

[3] When a global optimization procedure is preferred for θ in the case of highly nonlinear estimating equations, alternative algorithms are considered such as genetic or simplex algorithms.

3 Main generalizations of EL

The former empirical likelihood has been extended for particular estimating equation but also in its foundation. The likelihood which is maximized under constraints can be replaced with alternative functions. These generalization are selected among the many ones which have been proposed in the current literature.

3.1 MED and GEL

A generalization has resulted into a family of methods called generalized empirical likelihood (GEL) or minimum empirical discrepancy (MED) estimators. The following Table 1 summarizes some generalizations of the empirical likelihood approach, with other functions than a likelihood of the probability vector $\mathbf{w}$. In this generalization of EL named GEL,

Table 1: Estimator of type empirical likelihood.

	γ	$\phi(\mathbf{w})$	$\boldsymbol{\rho}(\boldsymbol{\zeta})$	$\boldsymbol{\tau}(\boldsymbol{\zeta})$
EL	-1	$ln\, nw$	$ln(1+\zeta)$	$(1+\zeta)^{-1}$
ET	0	$-nw\, ln\, nw$	$-\exp(\zeta)$	$-\exp(\zeta)$
CU	1	$-(nw)^2$	$-2^{-1}(1+\zeta)^2$	$-(1+\zeta)$
ECR	γ	$-\frac{(nw)^{\gamma+1}}{\gamma(\gamma+1)}$	$-\frac{(1+\gamma\zeta)^{\frac{\gamma+1}{\gamma}}}{\gamma+1}$	$-(1+\gamma\zeta)^{1/\gamma}$

with members EL, ET, CU and ECR, the former likelihood function is replaced with the new function:

$$r_{\text{GEL}}(\xi,\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log\rho(\xi'\phi_i(\theta))\,.$$

The related optimization problem leads to find out the solution for:

$$(\hat{\xi},\hat{\theta}) = \arg\max_\theta \inf_\xi \left\{r_{\text{GEL}}(\xi,\theta)\right\}\,.$$

Note that this problem is generally broken into two optimizations problems which are solved separately and iteratively by repeating until convergence: knowing θ the quantity ξ is estimated, then knowing ξ the quantity θ is estimated.

This problem is also the dual [38] to the direct problem in the MED approach when in $L(\mathbf{w})$ the logarithm function is replaced by one of the functions $\phi(\mathbf{w})$ in Table 1. The n weights are computed before looking for (ξ,θ) instead of directly estimate the later couple of parameters in the GEL approach. Note that this correspondence is obtained for the Cressie-Read (CR) discrepancies $\phi(\mathbf{w})$ as given in Table 1 for the more general case ECR. This is not always true because the MED approach does not always has a closed-form solution. For this case, the weights have a simple expression which insures their summation to one as and which can be written after resolving for the couple of parameters,

$$\hat{w}_i = \hat{w}_i(\hat{\xi},\hat{\theta}) = \frac{\tau(\hat{\xi}'\phi_i(\hat{\theta}))}{\sum_{i'}\tau(\hat{\xi}'\phi_i(\hat{\theta}))}\,.$$

Note that the same expression for the weigths is involved in the MED approach for a direct estimation. For instance as $\boldsymbol{\tau} = d\boldsymbol{\rho}/d\boldsymbol{\zeta}$, the case of EL is retrieved with the derivative of the logarithm as an inverse function. This can be justified as follows.

Sketch of the proof (see [39]) *In the exponential tilting EL, the optimization problem of the objective function under constraints leads to the maximization of:*

$$\begin{aligned} M(\theta) &= \textstyle\sum_i w_i \log w_i + \xi_{\mathbf{w}}(1-\sum_i w_i) \\ &\quad -\xi'(\textstyle\sum_i w_i\,\phi_i)\,. \end{aligned}$$

The partial derivatives of $M(\theta)$ with respect to w_i are set to zero such that:

$$\frac{\partial M(\theta,\xi_{\mathbf{w}},\xi)}{\partial w_i} = \log w_i + 1 - \xi_{\mathbf{w}} - \xi'\phi_i = 0_\ell\,.$$

By computing the partial derivative of $M(.)$ w.r.t. w_i, and as summing the resulting quantities for w_i depending on $\xi_{\mathbf{w}}$ equals to one, this leads a value for $\xi_{\mathbf{w}}$. Finally, one gets the following solution depending on the unknown θ, as:

$$w_i = w(z_i,\xi,\theta) = \frac{\exp(\xi'\phi_i)}{\sum_{i'}\exp(\xi'\phi_i)}\,,$$

which is the wanted expression introduced above. □

Note that computational algorithm for ET is discussed in [10] and [40]. On the contrary to empirical likelihood, the weights involve an exponential function and their summation to one is more obvious from their expression. A consequence is a more stable numerical behavior of the estimator but in the current literature it may perform more poorly in term of the efficiency, this explains why other alternative estimators have been proposed. A pitfall for any practical application of EL is its weakness against some misspecifications of the constraints. On the other hand, ET (exponential tilting) exhibits better bahaviors with finite variance in this situation. For these reasons, another estimator was defined for the best of the EL and ET as explained next subsection.

3.2 ETEL

The EL and ET estimators are combined in the exponentially tilted empirical likelihood (ETEL) estimator in order

to share the good properties of ET under misspecified models and similar higher-order properties than EL. Roughly speaking, ETEL is defined by 1) finding the expression for the weights with the entropic function as in the case ET -under the usual constraints- and 2) solving for θ under the knowledge of these weights with the logarithm function as in the case EL. The dual problem reduces to solve for:

$$\begin{aligned}\hat{\theta} &= \text{argmax}_\theta \left\{\tfrac{1}{n}\sum_{i=1}^n \log(n\, w_i(\xi,\theta)\right\} \\ &= \text{argmax}_\theta \left\{-\log(\tfrac{1}{n}\ \textstyle\sum_i \exp(\xi'(\phi_i - \bar{\phi}))\right\}\ .\end{aligned}$$

$$\begin{aligned}\hat{\xi} &= \text{argmax}_\xi \left\{-\tfrac{1}{n}\textstyle\sum_{i=1}^n \exp(\xi'\phi_i(\theta)\right\} \\ &= \text{argmax}_\xi \left\{-\log(\tfrac{1}{n}\textstyle\sum_{i=1}^n \exp(\xi'\phi_i(\theta))\right\} \\ &= \text{argmax}_\xi \left\{\tfrac{1}{n}\textstyle\sum_{i=1}^n \log(n\, w_i(\xi,\theta)\exp(-\xi'\phi_i(\theta))\right\} \\ &= \text{argmax}_\xi \left\{\tfrac{1}{n}\textstyle\sum_{i=1}^n \log(n\, w_i(\xi,\theta)) - \xi\bar{\phi}\right\}\end{aligned}$$

Here $\bar{\phi} = \sum_i \phi_i/n$ is a average from the functional constraints, and $w_i(\xi,\theta)$ is written as in the previous subsection with a normalized exponential function (multiplied by minux one) for $\tau(.)$. The expression of the maximization w.r.t. ξ as a function of the weights is new to our knowledge. Note that the justification is similar to maximum likelihood maximization as $d\log(f) = df/f$, the maximum remains the same. Let denote the (pseudo-)likelihood,

$$L_{\text{ETEL}}(\xi,\theta) = \prod_{i=1}^n (n\, w_i(\xi,\theta))\,.$$

The optimization problem above can be rewritten as,

$$\begin{aligned}\hat{\theta} &= \text{argmax}_\theta \left\{\tfrac{1}{n}\log L_{\text{ETEL}}(\xi,\theta)\right\} \\ \hat{\xi} &= \text{argmax}_\xi \left\{\tfrac{1}{n}\log L_{\text{ETEL}}(\xi,\theta) - \xi'\bar{\phi}\right\}\ .\end{aligned}$$

On the contrary to EL, ET and ETEL do not require the constraints of positivity of $\xi'\phi_i$, which facilitates the implementation of the algorithm.

3.3 ETHD

More rencently, a new empirical likelihood is proposed in [41], for enhancing the robustness of empirical likelihood estimator against misspecified models with the help of the minimum Hellinger Distance (HD). Following the idea of ETEL, this new estimator named exponentially tilted Hellinger distance (ETHD) considers its first step which provides the ET estimator and its second step with a new distance. More precisely, this leads to solve for:

$$\begin{aligned}\hat{\xi} &= \text{argmax}_\xi \left\{-\tfrac{1}{n}\textstyle\sum_{i=1}^n \exp(\xi'\phi_i(\theta)\right\} \\ \hat{\theta} &= \text{argmax}_\theta \left\{2 - 2\tfrac{1}{n}\textstyle\sum_{i=1}^n \sqrt{w_i(\hat{\xi},\theta)}\right\}\ .\end{aligned}$$

Here, the weigths have the same form than in the ET estimator as the ratio of an exponential function and a sum of exponential ones. According to the authors this estimator may be more robust and they present several analytical justications through mathematical proofs. This version of empirical likelihood may be preferred to ET and EL for its enhanced robutness.

3.4 Selected variants and improvements

Recent researches aims at improving EL in order to make its inference more stable or to improve its first order statistics.

The adjusted empirical likelihood[42, 43] is able to improve the behavior of the estimator from the numerical point of view for the convergence towards a solution and for the running time, but also from the point of statistical point of view with better coverage probabilities. Moreover, the authors have shown that this variant has the same asymptotic properties as the unadjusted empirical likelihood. It is defined as follows. Let have a_n a positive constant such as a new sample unit with a new unknown weight w_{n+1} is associated in the estimating equation to:

$$\phi_{n+1} = -a_n\bar{\phi}\,.$$

The new estimating equation counts this new datum with a new weight to be estimate jointly with the n former ones from the previous sample. The rational of this method is to insure that zero is in the convex hull of the new set of functions ϕ_i with the new one more. A consequence is that the search for the weights has always a solution on the contrary to the usual empirical likelihood. The authors propose to use the algorithm 1 for the numerical estimation except that now the sample has $n+1$ units, and they also observe that the converge happens more quickly than without the adjustment in the presented simulations. Note that for mean estimation, this variant has been improved and called balanced augmented empirical likelihood [44] which adds two new sample units instead of just one. For a similar multidimensional problem, a penalized empirical likelihood [45] was proposed by relaxing the equality to zero of the estimating equation: a multi-dimensional vector replaces the null value such that its Euclidean norm is used for penalizing the empirical likelihood ratio.

Mathematical theories and formula for the improvement of the small sample properties of empirical likelihood can be found also in [46, 47, 48, 49, 50] which underlines that the exponential EL cannot be improved via these approaches. For high dimensional parameters [29], the theory of empirical

likelihood needs to be updated because the multiplier ξ has new properties.

3.5 Weighted robust estimators (WEL)

For bringing robutness to empirical likelihood, authors have added weights in the loss function in different ways as the pioneer research in [51, 52]. See also [53] as explained in next subsection for sampling theory. Note that the weights may be added for improving the properties of empirical likelihood. In [54, 55] the chosen weights are directly related to current usual robust methods in parametric likelihood with the following expression:

$$\omega_i = min\left[1, \left(\frac{b_0}{(z_i - m_z)^T S_z^{-1}(z_i - m_z)}\right)^{\gamma_\omega/2}\right].$$

This weight is a distance between z_i and m_z where m_z and S_z are respectively robust estimates of the location and scale of the z_i. Note that an alternative[4] with directly the estimating equation ϕ_i instead of z_i seems not have been studied to our knowledge.

3.6 Weighted survey estimators

In survey theory, a major technical tool to proceed before all is to introduce weights [56] which account for the sampling of the n selected units in the sample S from the large population $U = (u_{(1)}, \ldots, u_{(N)})$ of size N. Typically, the weights are denoted $p_i = Pr(u_{(i)} \in S)$, while their inverse values are $d_i = 1/p_i$ -aggregated in the vector $\mathbf{d}$- and their normalized versions are $\tilde{d}_i = d_i/(\sum_i d_i)$. One of the more widely studied question for studying data survey is how to estimate a quantity of interest for a variable $\mathcal{Y}$ such as a mean from a sample $(y_1, \cdots, y_n)$. Another variable $\mathcal{X}$ with observed values $(x_1, \cdots, x_n)$ is also available at the level of the sample but also to the whole population for at least its total mean. This kind of additional variable serves to select the sample or to adjust the estimates of the quantity of interest thanks to the external information it provides.

The family of methods we are interested here is called calibration[57, 58] with the idea that new weights can be found by altering the former ones d_i in order to insure that the corresponding sample weighted estimate of the mean for $\mathcal{X}$ equals the true mean from the population, denoted t_x. Note that some of the functions $\phi(.)$ in Table 1 can be also found among the distances between original weights and calibrated versions in Table 1. (page 378) of [57]. This leads to an estimate of the mean for $\mathcal{Y}$ by considering these new weights. In empirical likelihood the weights are coming from the maximization of the likelihood of the ratio. Several approaches have been proposed in the current literature via empirical likelihood. Pioneering research in the domain can be found by reading the overview [59]. Just after, it is presented several methods from the current literature on complex survey. A bayesian pseudo-empirical-likelihood was proposed in [60] and out of the scope here. A pseudo empirical likelihood was proposed in [2]. A profile pseudo-empirical log-likelihood function was proposed in [3]. The generalized pseudo empirical likelihood was proposed in [56]. The authors have proposed a weighted Kullback-Leibler distance based calibration estimator by minimizing an alternative criterion to the weighted likelihood,

$$EL(\mathbf{d}, \mathbf{w}) = -\sum_{i \in S} q_i^{-1} \left\{ d_i \log \frac{w_i}{d_i} - w_i + d_i \right\}.$$

Here the quantites q_i are fixed constants often equal to 1 even if in alternative values be interesting in certain cases as in [57] where the same criterion is named minimum entropy distance. For surveys, the constraints are $\sum_{i \in S} w_i = N$ if the population size is known and $\sum_{i \in S} w_i x_i = \hat{t}_x$, which as put in only one constraint by adding 1 to the vectors x_i. The derivative of $EL(\mathbf{w})$ completed with the lagragian multipliers and the constraints lead to the solution for the weights:

$$\hat{w}_i = \frac{d_i}{1 + q_i x_i' \hat{\xi}}.$$

Note that this expression is retrived by several authors. Here $\hat{\xi}$ is found by solving for the constraint on the extern variables $\sum_{i \in S} \hat{w}_i x_i = \hat{t}_x$. This leads to an expression for a mean estimator from $(y_1, \ldots, y_n)$, say $\hat{t}_y = \sum_{i \in S} \hat{w}_i y_i$ from the estimated weights. The authors also suggests that minimizing instead of $EL(\mathbf{d}, \mathbf{w})$ the alternative version $EL(\mathbf{w}, \mathbf{d})$ leads to other weights related to exponential tilting developped in [61]. They also proposed to use the quantities $q_i = \pi_i^{-1} - 1$ in order to improve the properties of the estimator as in [61]. An algorithm similar to Algorithm 1 is used in the simulations. In [62, 63], the authors have proposed new empirical likelihood confidence intervals for complex sampling designs when the weights are introduced at the level of the estimating equations instead of the likelihood.

[4] This may suggest an alternative weighting for decreasing the influence of outliers in the set of values ϕ_i by updating the previous weighting with a following expression, $\omega_i = min\,[1, (b_0/(\phi_i - m_\phi)^T S_\phi^{-1}(\phi_i - m_\phi))^{\gamma_\omega/2}]$. This weight is a distance between ϕ_i and m_ϕ where m_ϕ and S_ϕ are robust estimates for the location and scale of the values ϕ_i, hence alternative expressions are available from M-estimation for instance.

4 Conclusion and perspectives

Herein it is proposed a brief overview of the computational approaches in empirical likelihood with the definition of the weights, their estimation via the inference of the lagrangian multipliers and of the parameters in the estimating equations. The generalization of empirical likelihood, via diverse definitions of the objective functions invoved, is also presented. Examples of efficient modeling via empirical likelihood with their analytical justifications are numerous in many domains such as clustering for modeling mixture models [64, 65] as proposed in [66]. A clustering algorithm was proposed with $\phi(x_i, \theta_k) = x_i - \mu_k$, thus defined without any form of distributional hypothesis. From its comparison with the usual agglomerative (k-means, k-medoids) and hierarchical (average, complete, single, ward) alternatives, this algorithm[5] is shown relevant. A main perspective for empirical likelihood may remain the developement of fast, robust and stable procedures for the parameters estimation of the many estimating equations involved.

In the paper, it is presented an introduction of the numerical algorithms for inference of EL and several of it generalizations. In the next paragraphs, it is also presented several applications of the framework of empirical likelihood within several recent modeling approach from the current literature of parametric likelihood. The proposed approach may be seen as convenient variants for a robust and flexible modeling in order to facilitate the estimation of the parameters in practice as a perspective.

Surrogate bound: The maximization of the likelihood of the ETEL estimator leads to the optimization for softmax functions which are approximated in the current literature. For empirical likelihood, this reduces to bound each quantity w_i such that this results into a bound for the whole ratio: the resulting surrogate function is maximized instead of the original objective function. Note that this kind of approximation was shown successfull in many domains such as for small area estimation [69] for instance. Let denote $\rho(\psi_j) = \frac{1}{2\psi_j}[\frac{1}{1+e^{\psi_j}} - \frac{1}{2}]$. Then the bound is obtained via two consecutive bounds as stated in [70, 71]. A first step leads to a product of sigmoid functions while a second step is their quadratic function approximation, such as finally, this may induce for instance that, with κ a given vector,

$$\begin{aligned} C(\xi, \theta) &= \frac{1}{n}\sum_{i=1}^n \log(n\, w_i(\xi, \theta)) - \xi'\kappa \\ &\geq \frac{1}{n}\sum_{i=1}^n \Big\{ -\alpha - \sum_{j=1}^n \frac{\xi'\phi_j - \alpha - \psi_j}{2} \\ &\quad -\rho(\psi_j)[(\xi'\phi_j - \alpha)^2 - \psi_j^2] \\ &\quad -\log(1 + e^{\psi_j})\Big\} + \xi'(\bar{\phi} - \kappa) \\ &\quad +n \log n\,. \end{aligned}$$

This surrogate bound has several advantages on the original expression for the weights with an eventual variational bayesian inference. The quadratic function can be solved in one closed-form step for the multiplier ξ. Adding the constraints with inequalities for the x_i as in the EL case, may be more convenient. Similarly, the solution for the parameters can be easier to handle for the parameters θ, for regression function ϕ_i for instance or even for more nonlinear functions ϕ_i. This suggests also possible a closed-form solution for adding a prior on ξ or θ, for the inference [72, 73] of a bayesian version[6]. These approaches are also appealing when sample size increase, as robustness becomes even more important for the inference.

Robust divergences: For introducing further robutness in the inference procedure, it may be considered a variant for ETEL that is called here ETDV for ET with a robust divergence. Following recent researches, this leads to chose a more robust objective criterion than the Kullback-Leibler divergence (see also [76]) when the parameters are optimized in the ET variant. Candidate ones [77, 78] are among widely used $\alpha-$, $\beta-$, $\gamma-$, Rényi- divergences for instance. Some divergences are presented in the Table 2 but other ones [79] are possible.

This approach may be more relevant for larger sample sizes when the weights are closer to $\frac{1}{n}$ except for the outliers.

[5] It is supposed x_i univariate and the following model defined for ℓ clusters as follows:

$$L(\mathbf{w}) = \prod_{i=1}^n \prod_{k=1}^{\ell} \{\pi_k w_{ik}(\xi_k, \theta_k)\}^{z_{ik}}\,,$$

where $z_{ik} \in \{0, 1\}$, $w_{ik} = w(z_{ik}, \xi, \theta)$ with n_k instead of n, and $\phi(x_i, \theta_k) = (x_i - \mu_k)$. The estimating equations may come from the first moments of a Gaussian distribution [13] with mean μ_k and variance σ_k while π_k are for the mixing components. For the inference, a typical classifying EM algorithm [67, 68] is possible in two steps.

[6] The bayesian EL [72] has been proposed in the literature. Its author shows that asymptotically this methods leads to similar estimators than the non bayesian methods, BEL for EL and BETEL for ETEL such as the influence of the prior may vanishes for large n. The B(ET)EL posterior density is defined as being proportional to the product of the prior $\pi(.)$ for θ and the (ET)EL, $\prod_i^n w_{i;\mathrm{BETEL}} \propto \pi(\theta)\prod_i^n w_i$, see also more recently [74]. Typically, bayesian procedures are involved for the inference of the unknown quantities. For reducing the impact of outliers on the estimation of population moments, a new robust BETEL inferential methodology is recently proposed [75] in the model RBETEL where the authors introduces a divergence for the justification of the non robust former method.

Table 2: Example of robust divergences.

Name	Expression
α-.	$D_\alpha\left(\mathbf{w}\|\|\frac{\mathbf{1}}{n}\right) = \frac{\sum_i w_i^\alpha \frac{1}{n^{1+\alpha}} - \alpha w_i + (\alpha-1)\frac{1}{n}}{\alpha(\alpha-1)}$
β-.	$D_\beta\left(\mathbf{w}\|\|\frac{\mathbf{1}}{n}\right) = \frac{\sum_i w_i^{\beta+1} + \beta \frac{1}{n^{1+\beta}} - (\alpha+1) w_i \frac{1}{n^\beta}}{\beta(\beta+1)}$
$\alpha-\beta$-.	$D_\alpha^\beta\left(\mathbf{w}\|\|\frac{\mathbf{1}}{n}\right) = -\frac{\left(\sum_i w_i^\alpha \frac{1}{n^\beta} - \frac{\alpha}{\alpha+\beta} w_i^{\alpha+\beta} + \frac{\beta}{\alpha+\beta}\frac{1}{n^{\alpha+\beta}}\right)}{\alpha\beta}$
γ-.	$D_\gamma\left(\mathbf{w}\|\|\frac{\mathbf{1}}{n}\right) = \frac{1}{\gamma(\gamma+1)} \log \frac{\left(\sum_i w_i^{\gamma+1}\right)\left(\sum_i \frac{1}{n^{1+\gamma}}\right)^\gamma}{\left(\sum_i w_i \frac{1}{n^\gamma}\right)^{\gamma+1}}$
Rényi .	$D_\rho\left(\mathbf{w}\|\|\frac{\mathbf{1}}{n}\right) = \frac{1}{\rho-1} \log\left(\sum_i w_i^\rho \frac{1}{n^{1-\rho}}\right)$

Note that the criterion of ETHD is a particular case with the Hellinger distance because it is part of the α-divergence - when $\alpha = 0.5$ - with the theoretical justifications from their authors. A modified empirical likelihood requires not only its own numerical procedure for the parameter inference but also new properties in order to eventually improve the state of art.

Additive noise: Robustness is usually added to the model itself before inference via empirical likelihood. Some mathematical research focus on the theory of robustness under misspecification. This has motivated recent variants of the empirical likelihood such as ETEL and ETHD which construct an estimator in two step: finding the weights with also the lagrangian multipliers via ET and infer the parameters from the estimating equations via another model.

As an alternative to weighting, by following [80] for reducing the impact from the behavior of the outliers which induce locally a mistaken modeling, it can be added a noise to the estimating equations such as ϕ_i becomes[7] as follows:

$$\tilde{\phi}_i = \phi_i + e_i \,.$$

Here e_i is a centered white noise, eventually a t-Student law for better reflecting the effects coming only from few outliers. This kind of approach has been shown successful in likelihood methods for variational bayesian learning. The optimization problem is as previously except that the quantities e_i are also to be found by adding a penalization with their squares. Here, for the likelihood cases, this can be seen as if the original noise in the model (leading to the estimating equations) is a function of the parameter such as at the derivative step a quantity remains, otherwise, this is just a way to manage the too large or too small values of ϕ_i coming from the misspecification of a few observations.

[7] Or even $w_i = w(z_i, \xi, \xi_\gamma, \theta) = \frac{1}{n}\frac{1}{1+e_i+\xi'\phi(z_i,\theta)+\xi'_\gamma\phi(z_i)}$ when the quantity $\xi'\phi(z_i,\theta) + \xi'_\gamma\phi(z_i)$ does not respect the constraints for some sample unit.

References

[1] A. Owen, Empirical Likelihood, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press, 2001.

[2] J. Chen, R. R. Sitter, C. Wu, Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys, Biometrika 89 (1) (2002) 230–237.

[3] C. Wu, J. N. K. Rao, Pseudo-empirical likelihood ratio confidence intervals for complex surveys, Can. J. Statist. 34 (2006) 359–375.

[4] N. Reid, Likelihood inference in complex settings, The Canadian Journal of Statistics 40 (4) (2012) 731–744.

[5] I. J. Myung, Tutorial on maximum likelihood estimation, J. Math. Psychol. 47 (1) (2003) 90–100.

[6] A. Henningsen, O. Toomet, maxlik: A package for maximum likelihood estimation in r, Comput Stat 26 (2011) 443–458.

[7] J. Qin, J. Lawless, Empirical likelihood and general estimating equations, The Annals of Statistics 22 (1) (1994) 300–325.

[8] A. Owen, Empirical likelihood ratio confidence regions, Ann. Statist. 18 (1) (1990) 90–120.

[9] A. Owen, Empirical likelihood for linear models, Ann. Statist. 19 (4) (1991) 1725–1747.

[10] G. W. Imbens, Generalized method of moments and empirical likelihood, Journal of Business & Economic Statistics 20 (4) (2002) 493–506.

[11] C. Wu, Some algorithmic aspects of the empirical likelihood method in survey sampling, Statistica Sinica 14 (2004) 1057–1067.

[12] C. Wu, Algorithms and r codes for the pseudo empirical likelihood method in survey sampling, Survey Methodology 31 (2005) 239–243.

[13] P. Chaussé, Computing generalized method of moments and generalized empirical likelihood with r, Journal of Statistical Software 34 (11) (2010) 1–35.

[14] M. Li, L. Peng, Y. Qi, Reduce computation in profile empirical likelihood method, The Canadian Journal of Statistics / La Revue Canadienne de Statistique 39 (2) (2011) 370–384.

[15] D. Yang, D. S. Small, An r package and a study of methods for computing empirical likelihood, Journal of Statistical Computation and Simulation 83 (7) (2013) 1363–1372.

[16] C. Y. Tang, T. T. Wu, Nested coordinate descent algorithms for empirical likelihood, Journal of Statistical Computation and Simulation 84 (9) (2014) 1917–1930.

[17] Y. Zhao, A. Vexler, A. Hutson, X. Chen, A statistical software procedure for exact parametric and nonparametric likelihood-ratio tests for two-sample comparisons, Communications in Statistics - Simulation and Computation 46 (4) (2017) 2829–2841.

[18] H. Lam, H. Qian, The empirical likelihood approach to simulation input uncertainty, in: Proceedings of the 2016 Winter Simulation Conference, 2016.

[19] P. Hall, B. L. Scala, Methodology and algorithms of empirical likelihood, International Statistical Review / Revue Internationale de Statistique 58 (2) (1990) 109–127.

[20] Y. Kitamura, Empirical likelihood methods in econometrics: Theory and practice, in: W. K. N. R. Blundell, T. Personn (Eds.), Advances in Economics and Econometrics: Ninth World Congress of the Econometric Society, Cambridge University Press, 2007.

[21] S. X. Chen, I. Van Keilegom, A review on empirical likelihood methods for regression, TEST 18 (3) (2009) 415–447.

[22] P. M. Parente, R. J. Smith, Recent developments in empirical likelihood and related methods, Annual Review of Economics 6 (1) (2014) 77–102.

[23] S. Chaudhuri, M. S. Handcock, M. S. Rendall, Generalized linear models incorporating population level information: An empirical-likelihood-based approach, Journal of the Royal Statistical Society. Series B (Statistical Methodology) 70 (2) (2008) 311–328.

[24] S. Chaudhuri, M. S. Handcock, M. S. Rendall, A conditional empirical likelihood approach to combine sampling design and population level information, Tech. rep., National University of Singapore (2010).

[25] J. K. Hellerstein, G. W. Imbens, Imposing moment restrictions from auxiliary data by weighting, Rev. Econ. Statist. 81 (1999) 1–14.

[26] M. Luersen, R. Le Riche, F. Guyon, A constrained, globalized, and bounded nelder-mead method for engineering optimization, Structural and Multidisciplinary Optimization 27 (1) (2004) 43–54.

[27] E. M. M. Lopez, I. V. Keilegom, N. Veraverbeke, Empirical likelihood for non-smooth criterion functions, ArXiv e-prints.

[28] A. B. Owen, Self-concordance for empirical likelihood, The Canadian Journal of Statistics / La Revue Canadienne de Statistique 41 (3) (2013) 387–397.

[29] J. Chang, C. Y. Tang, T. T. Wu, A new scope of penalized empirical likelihood with high-dimensional estimating equations, ArXiv e-printsarXiv:1704.00566.

[30] J. B. Lasserre, Why the logarithmic barrier function in convex and linear programming?, Operations Research Letters 27 (4) (2000) 149 – 152.

[31] C.-J. Lin, R. C. Weng, S. S. Keerthi, Trust region newton method for logistic regression, J. Mach. Learn. Res. 9 (2008) 627–650.

[32] L. Laporte, R. Flamary, S. Canu, S. Déjean, J. Mothe, Nonconvex regularizations for feature selection in ranking with sparse svm, IEEE Transactions on Neural Networks and Learning Systems 25 (6) (2014) 1118–1130.

[33] S. Lee, S. Kwon, Y. Kim, A modified local quadratic approximation algorithm for penalized optimization problems.

[34] T. Tsiligkaridis, E. Marcheret, V. Goel, A difference of convex functions approach to large-scale log-linear model estimation, IEEE Transactions on Audio, Speech, and Language Processing 21 (11) (2013) 2255–2266.

[35] F. Espuny-Pujol, K. Morrissey, P. Williamson, A global optimisation approach to range-restricted survey calibration, Stat Comput (2017) 1–13.

[36] J. Ma, Multiplicative algorithms for maximum penalized likelihood inversion with non negative constraints and generalized error distributions, Communications in Statistics - Theory and Methods 35 (5) (2006) 831–848.

[37] L. van der Maaten, G. Hinton, Visualizing Data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605.

[38] T. Nishimura, F. Komaki, The information geometric structure of generalized empirical likelihood estimators, Communications in Statistics - Theory and Methods 37 (12) (2008) 1867–1879.

[39] S. M. Schennach, Point estimation with exponentially tilted empirical likelihood, The Annals of Statistics 35 (2) (2007) 634–672.

[40] H. Zhu, H. Zhou, J. Chen, Y. Li, J. Lieberman, M. Styner, Adjusted exponentially tilted likelihood with applications to brain morphology, Biometrics 65 (3) (2009) 919–927.

[41] B. Antoine, P. Dovonon, Robust estimation with exponentially tilted hellinger distance (February 2016).

[42] J. Chen, A. M. Variyath, B. Abraham, Adjusted empirical likelihood and its properties, Journal of Computational and Graphical Statistics 17 (2) (2008) 426–443.

[43] J. Chen, Y. Huang, Finite-sample properties of the adjusted empirical likelihood, Journal of Nonparametric Statistics 25 (1) (2013) 147–159.

[44] M. K. Nguyen, S. Phelps, W. L. Ng, Simulation based calibration using extended balanced augmented empirical likelihood, Stat Comput 25 (2015) 1093–1112.

[45] F. Bartolucci, A penalized version of the empirical likelihood ratio for the population mean, Stat Probab Lett 77 (2007) 104–110.

[46] T. DiCiccio, P. Hall, J. Romano, Empirical likelihood is bartlett-correctable, Ann. Statist. 19 (2) (1991) 1053–1061.

[47] B.-Y. Jing, A. T. A. Wood, Exponential empirical likelihood is not bartlett correctable, The Annals of Statistics 24 (1) (1996) 365–369.

[48] M. Tsao, F. Wu, Empirical likelihood on the full parameter space, Ann. Statist. 41 (4) (2013) 2176–2196.

[49] S. X. Chen, H. Cui, On bartlett correction of empirical likelihood in the presence of nuisance parameters, Biometrika 93 (1) (2006) 215–220.

[50] L. Camponovo, T. Otsu, On bartlett correctability of empirical likelihood in generalized power divergence family, Statistics & Probability Letters 86 (2014) 38–43.

[51] N. Glenn, Y. Zhao, Weighted empirical likelihood estimates and their robustness properties, CSDA 51 (10) (2007) 5130–5141.

[52] H. Ding, K. Lam, Weighted empirical likelihood estimator for vector multiplicative error model, Journal of Forecasting 32 (7) (2013) 613–627.

[53] C. Wu, Weighted empirical likelihood inference, Statistics & Probability Letters 66 (1) (2004) 67–79.

[54] G. Qin, Y. Bai, Z. Zhu, Robust empirical likelihood inference for longitudinal data, Statistics and Probability Letters 79 (2009) 2101–2108.

[55] G. Qin, Y. Bai, Z. Zhu, Robust empirical likelihood inference for generalized partial linear models with longitudinal data, Journal of Multivariate Analysis 105 (2012) 32–44.

[56] Z. Tan, C. Wu, Generalized pseudo empirical likelihood inferences for complex surveys, Canadian Journal of Statistics 43 (1) (2015) 1–17.

[57] J.-C. Deville, C.-E. Sarndal, Calibration estimators in survey sampling, JASA 87 (418) (1992) 376–382.

[58] C. Wu, W. W. Lu, Calibration weighting methods for complex surveys, International Statistical Review.

[59] J. N. K. Rao, Empirical likelihood methods for sample survey data: An overview, Austrian journal of statistics 35 (2-3) (2006) 191–196.

[60] J. N. K. Rao, C. Wu, Bayesian pseudo-empirical-likelihood intervals for complex surveys, J. R. Statist. Soc. B 72 (4) (2010) 533–544.

[61] J. K. Kim, Calibration estimation using exponential tilting in sample surveys, Survey Methodology 36 (2010) 145–155.

[62] Y. Berger, O. Torres, Empirical likelihood confidence intervals for complex sampling designs, Journal of the Royal Statistical Society: Series B (Statistical Methodology) (2016) 1–23.

[63] Y. Berger, Wiley StatsRef: Statistics Reference Online, Wiley, 2017, Ch. Empirical likelihood approaches under complex sampling designs, doi:10.1002/9781118445112.

[64] J. A. Bilmes, A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models, Tech. rep., ICSI, U.C. Berkeley (1998).

[65] G. J. McLachlan, D. Peel, Finite mixture models, Wiley Series in Probability and Statistics, New York, 2000.

[66] V. Melnykov, G. Shen, Clustering through empirical likelihood ratio, CSDA 62 (2013) 1–10.

[67] A. Dempster, N. Laird, D. Rubin, Maximum-likelihood from incomplete data via the EM algorithm, J. Royal Statist. Soc. Ser. B., 39 (1977) 1–38.

[68] A. Juan, E. Vidal, On the use of bernoulli mixture models for text classification, Pattern Recognition 35 (12) (2002) 2705 – 2710.

[69] F. Rijmen, J. Vomlel, Assessing the performance of variational methods for mixed logistic regression models, Journal of Statistical Computation and Simulation 78 (8) (2008) 765–779.

[70] G. Bouchard, Efficient bounds for the softmax function and applications to approximate inference in hybrid models.

[71] N. Depraetere, M. Vandebroek, A comparison of variational approximations for fast inference in mixed logit models, Computational Statistics 32 (1) (2017) 93–125.

[72] N. A. Lazar, Bayesian empirical likelihood, Biometrika 90 (2) (2003) 319–326.

[73] S. Chaudhuri, D. Mondal, T. Yin, Hamiltonian monte carlo sampling in bayesian empirical likelihood computation, Journal of the Royal Statistical Society Series B 79 (1) (2017) 293–320.

[74] A. Yiu, R. J. B. Goudie, B. D. M. Tom, Inference under unequal probability sampling with the Bayesian exponentially tilted empirical likelihood, Biometrika 107 (4) (2020) 857–873.

[75] Z. Liu, C. S. Forbes, H. M. Anderson, A Robust Bayesian Exponentially Tilted Empirical Likelihood Method, ArXiv e-printsarXiv:1801.00243.

[76] C. Wu, R. Zhang, An information-theoretic approach to the effective usage of auxiliary information from survey data, AISM 58 (2006) 499–509.

[77] M. Basseville, Review: Divergence measures for statistical data processing-an annotated bibliography, Signal Process. 93 (4) (2013) 621–633.

[78] O. Dikmen, Z. Yang, E. Oja, Learning the information divergence, IEEE Trans. Pattern Anal. Mach. Intell. 37 (7) (2015) 1442–1454.

[79] A. Cichocki, S. Cruces, S. Amari, Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization, Entropy 13 (2011) 134–170.

[80] C. Wang, D. M. Blei, A General Method for Robust Bayesian Modeling, ArXiv e-printsarXiv:1510.05078.