Reference Priors for Empirical Likelihoods

Bertrand Clarke* Ao Yuan †

Abstract

Estimators based on side information from constraints of the form $E[g(X, \theta)] = 0$, where θ is a finite-dimensional parameter are popular in many settings. This is so because, in contrast to likelihood based estimators, they do not require full specification of the distribution of X. Empirical likelihoods are one class of such constrained optimization procedures that have been well-studied from a Frequentist standpoint, but have only recently been used for Bayesian analysis.

Here, we derive asymptotic expansions for the distance between a prior and the posterior it generates when applied to an empirical likelihood. By optimizing the leading prior-dependent term in these expansions we identify reference priors that can be used for analysis directly or for comparison to other priors. We find that the reference prior under the relative entropy and Hellinger distances are reciprocal and are based on a transformation of the expected outer product of the constraint function. The reference prior under the Chi-square distance is only exhibited as a solution to a calculus of variations problem.

1 Introduction

Since Owen (1988), Owen (1990), and Owen (1991), empirical likelihood (EL) techniques have gained popularity largely because they incorporate information for parameter estimation into a non-parametric context by constrained optimization. Despite extensive use in

*Department of Medicine, Center for Computational Sciences, and Department of Epidemiology and Public Health, University of Miami, 1120 NW 14th Street, CRB 611 (C-213), Miami, FL, 33136. Email: bclarke2@med.miami.edu Tel: 305-2435457. Fax: 305-243- 9304

†Statistical Genetics and Bioinformatics Unit, National Human Genome Center, Howard University, 2216 Sixth Street, N.W., Suite 206, Washington, DC 20059. Email: ayuan@howard.edu Tel: 202-806-4361. Fax: 202-265-0871. Yuan's works is partly supported by the National Center for Research Resources by NIH grant 2G12RR003048.

the Frequentist context, EL has only recently come into use in Bayesian analysis. Lazar (2003) observed that the properties of EL are in many respects similar to those of parametric likelihoods and proposed ways they could be used in Bayesian inference. She presented several simulations under different conditions to show the effect of prior selection in ELs was much the same as in independence likelihoods. Further similarities between ELs and parametric likelihoods have been delineated in Yuan et al. (2009). The implication of this is that reference priors for ELs may behave similarly to the way reference priors in independence likelihoods do. Specifically, they may give slightly narrower credibility sets than the normal priors with large variances as studied in Lazar (2003). In econometrics, Moon and Schorfheide (2004) used EL's for a Bayesian analysis by choosing priors that put most of their mass on parameter values for which the moment constraint was approximately satisfied. Recently, Grendar and Judge (2009) established that EL's can be regarded as a posterior mode in an asymptotic sense.

Here, our main contribution is the identification of reference priors for empirical likelihood. This is important because, in principle, once a model and prior have been chosen, the posterior is determined and Bayesian analysis can proceed computationally. Indeed, automating prior selection – regardless of whether the resulting priors are used to form credibility sets – can help with posterior exploration. We recall that reference priors are merely one class of objective priors, see Ghosh (2009) for a recent survey.

In the next section we briefly review the formulation of empirical likelihoods. Then, in Section 3, we review the concept of reference priors for IID likelihoods, set up the corresponding optimization problem for the EL, and quote a result that will help us solve it. In section 4, we state our main results giving asymptotic expansions for three distances between priors and posteriors obtained from ELs and identify the reference priors they generate. In Section 5, we discuss the implications of our work. Technical details are relegated to Appendices.

2 Empirical Likelihood

The basic formulation of EL as given by Owen (1988) can be expressed as in Qin and Lawless (1994) and stated as follows. Let $X^n = (X_1, ..., X_n)$ be IID d-dimensional random vectors with unknown distribution function F and suppose the q-dimensional parameter

$\theta = (\theta_1, \ldots, \theta_q)'$ is a functional value of F, i.e., there is a function T so that $T(F) = \theta$. Write $x^n = (x_1, \ldots, x_n)$ to denote outcomes of X^n and x to denote outcomes of an individual random variable X. Suppose that additional information linking θ and F is available from a set of functions $g(x, \theta) = (g_1(x, \theta), ..., g_r(x, \theta))'$ where $r \geq q$ and $E[g(X, \theta)] = 0$. The expectation is taken in the distribution $F = F_T$ taken to be true and it is assumed that the true θ, θ_T satisfies $T(F_T) = \theta_T$. Indeed, here we assume that θ is identifiable with respect to g, i.e., $\theta^* \neq \theta$ implies $E[g(X, \theta^*)] \neq 0$.

An expression for the EL can be given as follows. Let F be a distribution function varying over a class and consider the likelihood

$$L(F) = \prod_{i=1}^{n} F(\{x_i\}). \tag{2.1}$$

Now, write $w_i = F(\{x_i\})$ and $w = (w_1, ..., w_n)$. The EL is subject to the auxilliary information constraints from g and achieves

$$\max_w \prod_{i=1}^{n} w_i \quad \text{subject to} \quad \sum_{i=1}^{n} w_i = 1 \quad \text{and} \quad \sum_{i=1}^{n} w_i g(x_i, \theta) = 0.$$

Let $t = (t_1, ..., t_r)'$ be the Lagrange multipliers corresponding to the constriant with $g(x, \theta)$, then one can derive

$$w_i = \frac{1}{n} \frac{1}{1 + t'g(x_i, \theta)}$$

and that $t = t_n(x_1, ..., x_n, \theta)$ is determined by

$$\sum_{i=1}^{n} \frac{g(x_i, \theta)}{1 + t'g(x_i, \theta)} = 0. \tag{2.2}$$

Note that $t = 0$ satisfies (2.2), but then $w_i = 1/n$ and the side information from g does not enter the EL. So, we henceforth require $t \neq 0$ to avoid trivality. Thus, the empirical likelihood assumes the form

$$p_\theta^n = p(x^n|\theta) = \prod_{i=1}^{n} \frac{1}{n} \frac{1}{1 + t'g(x_i, \theta)} = \prod_{i=1}^{n} w_i. \tag{2.3}$$

Note that even though the data is IID F, (2.3) does not in general factor into a product of terms each depending on only one of the x_is. Thus, ELs are not in general independence likelihoods (unlike (2.1)) even though they may be regarded as identical. As such they are a generalization of IID to permit a dependence structure induced by the constraint. Indeed, the data enter the constraint symmetrically so we expect that they will remain symmetric

in the empirical likelihood itself. Because of this dependence, it is difficult to assign priors to ELs and so Bayesian analysis has been limited. Our main contribution is the extension of objective Bayes methods to the EL context by deriving reference priors for them.

3 Reference Priors

In a pair of seminal papers, Shannon (1948a) and Shannon (1948b) gave an outline of the general theory of communication. One of the basic ideas was to reinterpret the conditional density given a parameter, or likelihood, as an 'information theoretic channel'. The idea is that θ is now a message drawn from a source distribution of messages, say Π with density π, and the sender wants to send the value θ to a collection of receivers. The receivers, however, do not receive θ exactly. Each receiver for $i = 1, \ldots, n$ receives a noisy version of θ, say x_i, from which they want to decode θ. The relationship between the θ sent and the x received is given by $p(x|\theta)$; the difference between a channel and likelihood is that the channel is a conditional density that will be used repeatedly (with both arguments redrawn) whereas a likelihood is a function of θ for fixed x^n . Now, assume each of the n receivers receives an x_i independently of the rest, but they pool their x_is to decode θ. If this process occurs many times, Shannon showed the rate of information transmission is

$$I(\Theta; X^n) = \int \int \pi(\theta) p(x^n|\theta) \ln \frac{p(x^n|\theta)}{m(x^n)} \mu(dx^n) \mu(d\theta), \tag{3.4}$$

(in nats per symbol) where μ generically denotes a dominating measure for its argument. The quantity in (3.4) is the (Shannon) mutual information. The natural question is how large it can be. This is answered by maximizing over W to find the maximal rate, the capacity of the channel $p(\cdot|\cdot)$. The result is

$$\Pi_{cap}(\cdot) = \arg\max_{\Pi} \ I(\Theta; X^n),$$

the capacity achieving source distribution. Asymptotically in n, Ibragimov and Hasminsky (1973) showed Π_{cap} was Jeffreys prior for regular finite-dimensional parametric families.

Bernardo (1979) wrote

$$I(\Theta, X^n) = E_m D(\pi(\cdot) \| \pi(\cdot|X^n)),$$

where, for densities p and q with respect to a common dominating measure, the relative entropy is

$$D(p||q) = \int p(x) \ln \frac{p(x)}{q(x)} \mu(dx).$$

That is, the capacity achieving source distribution is the prior that makes the asymptotic distance between a prior and its corresponding posterior as far apart as possible, on average, in relative entropy. Bernardo (1979) also called Π_{cap} a reference prior on the grounds that it could be used as a prior, or more typically, used as a way to assess the amount of information in a subjective prior.

We comment that reference prior results are asymptotic in n and we assume this without further comment apart from noting that reference priors obtained for fixed n are usually discrete, see Berger et al. (1991). Even so, Zhang (1994) provides a convergence result ensuring the discrete priors converge to Jeffreys prior for many regular parametric families.

Berger and Bernardo (1989) examined a conditional form of the Shannon mutual information to identify

$$\begin{aligned} &\arg \max_{\Pi} I(\Theta; X^n|\Psi) \\ = \quad &\arg \max_{\Pi} \int \pi(\theta|\psi) p(x^n|\theta,\psi) \ln \frac{p(x^n|\theta,\psi)}{\int p(x^n|\theta,\psi)\pi(\theta|\psi)\mu(d\theta)} \mu(dx^n)\mu(d\theta)\pi(\psi)\mu(d\psi), \end{aligned}$$

where $(\Theta, \Psi) = (\Theta_1, \ldots, \Theta_q, \Psi_1, \ldots, \Psi_\ell)$ and Ψ is a nuisance parameter. A proof for regular finite-dimensional families can be found in Ghosh and Mukerjee (1992). Further treatment of the multiparameter case can be found in Berger and Bernardo (1991), Berger and Bernardo (1992a), and Berger and Bernardo (1992b). Sun and Berger (1998) examined conditional mutual information further and Clarke and Yuan (2004) gave a complete treatment.

Of recent interest is the work done by Ghosh et al. (2009) and Liu and Ghosh (2009) to obtain reference priors under alternative measures of distance. They establish that Jeffreys prior is the reference prior for almost all members of the power divergence family. The exception is the Chi-square distance for which the prior turns out to be proportional to the fourth root of the determinant of the Fisher information.

To be precise about the quantities we examine for the EL, write

$$m_n(x^n) = m(x^n) = \int p(x^n|\theta)\pi(\theta)d\theta,$$

where $p(x^n|\theta)$ is an in (2.3) , giving posterior

$$\pi(\theta|x^n) = \frac{\pi(\theta)p(x^n|\theta)}{m(x^n)}.$$

Then, the relative entropy between $\pi(\theta|x^n)$ and $\pi(\theta)$ is

$$D(\pi(\cdot|x^n)||\pi(\cdot)) = \int \pi(\theta|x^n) \log \frac{\pi(\theta|x^n)}{\pi(\theta)} d\theta;$$

the Hellinger distance between $\pi(\theta|x^n)$ and $\pi(\theta)$ is

$$H(\pi(\cdot|x^n), \pi(\cdot)) = \int (\sqrt{\pi(\theta)} - \sqrt{\pi(\theta|x^n)})^2 d\theta;$$

and the Chi-squared distance between $\pi(\theta|x^n)$ and $\pi(\theta)$ is

$$\chi^2(\pi(\cdot|x^n), \pi(\cdot)) = \int \frac{(\pi(\theta|x^n) - \pi(\theta))^2}{\pi(\theta)} d\theta.$$

Here, we examine the expectation of the above three quantities namely

$$E_{m_n} D(\pi(\cdot|x^n)||\pi(\cdot)) \quad E_{m_n} H(\pi(\cdot|x^n), \pi(\cdot)), \quad \text{and} \quad E_{m_n} \chi^2(\pi(\cdot|x^n), \pi(\cdot)).$$

These three distance measures have interpretations that may make them more or less useful in a given setting. The relative entropy occurs in probabilistic coding theory and usually represents an amount of information (in nats). The Chi-square distance is familiar from goodness-of-fit testing, see Clarke and Sun (1997), see also Herve (2007). The Hellinger distance originates from geometry in which the square root converts a great circle on the unit sphere to a line segment in a plane. It can be verified that as distances, $\chi^2(p, q) \geq D(p||q) \geq H(p, q)$ for any two densities p, q for which they are defined.

Observe that $m(x^n)$ is the Bayes action for estimating P_θ under relative entropy i.e.,

$$m(x^n) = \arg\min_Q \int w(\theta) D(P_\theta^n||Q) \mu(d\theta).$$

and the chain rule for relative entropy gives

$$D(P_\theta^n||M_n) = \sum_{k=1}^{n} E_m D(P_\theta||M_k(\cdot|X^{k-1})). \tag{3.5}$$

However, under Hellinger distance the Bayes action for estimating of P_θ is

$$m_H(x^n) = \left(\int w(\theta) p(x^n|\theta)^{1/2} \mu(d\theta) \right)^2 = \arg\min_Q \int w(\theta) H(P_\theta^n||Q) \mu(d\theta),$$

and under Chi-square distance the Bayes action for estimating P_θ is

$$m_{\chi^2}(x^n) = \int w(\theta) p(x^n|\theta)^2 \mu(d\theta) = \arg\min_Q \int w(\theta) \chi^2(P_\theta^n||Q) \mu(d\theta).$$

It is seen that neither is a probability (unless additional constraints are imposed) and neither satisfies an additive risk condition like (3.5). This means that the reference priors under Hellinger or Chi-square distance are not for the Bayes action and so need not be least favorable. However, they do provide priors maximally changed on average by the data.

Note that, to date, almost all reference prior work has been in the regular parametric family context. However, there are cases, such as EL, in which we do not have a well-defined IID parametric likelihood. Indeed, as can be seen from (2.3), the EL is stationary but not independent. However, the stationarity is close enough to independence that MLEs are consistent and Laws of Large Numbers and Central Limit Theorems hold.

To see this more formally, define the following notation. Let θ be the 'true' parameter value for the observed data and assume θ is in an open set whose closure is compact. Write $l_i(\theta) = \log w_i(\theta)$ with first derivative denoted $l_i^{(1)}(\theta) = \partial l_i(\theta)/\partial\theta$ and second derivative denoted $l_i^{(2)}(\theta) = \partial l_i(\theta)/[\partial\theta\partial\theta']$. Next, consider the following regularity conditions.

- R1: The constraint function has bounded moments, i.e., $E||g(X,\theta)||^\alpha < \infty$ for some $\alpha > 2$.
- R2: The outer product matrix $\Omega = E[g(X,\theta)g'(X,\theta)]$ is positive definite.
- R3: The Jacobian matrix $D = E[\partial g(X,\theta)/\partial\theta]$ is of rank r.
- R4: The norms $||g(x,\theta)||$ and $||g'(x,\theta)g(x,\theta)||$ are bounded by an integrable function function $G(x)$, in each neighborhood of θ.
- R5: The prior $\pi(\cdot)$ is continuous and the matrix $\Lambda(\theta) = D'(\theta)\Omega^{-1}(\theta)D(\theta)$ is invertible.
- R6: The prior $\pi(\cdot)$ and the $l_i^{(2)}(\cdot)$ for $i = 1, \ldots, n$ are bounded.

Denoting the maximum empirical likelihood estimate of θ by $\hat{\theta}_n = \arg\sup_\theta \log p(x^n|\theta)$, we have the following asymptotic results which parallel the corresponding results for regular IID likelihoods.

Theorem 0. *Assume R1–R4. Then $\hat{\theta}$ is consistent and asymptotically normal with asymptotic variance matrix $\Lambda^{-1}(\theta)$. That is,*

$$\hat{\theta}_n \to \theta \quad a.s. \quad \text{and} \quad \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, \Lambda^{-1}).$$

Proof: See Yuan et al. (2009).

4 Relative Entropy Reference Priors

Equipped with the EL setting of Section 2 and the reference prior formulation of Section 3, we can now state the first of our main results.

Theorem 1. *Assume R1–R6. Then*

$$E_{m_n} D(\pi(\cdot|x^n)||\pi(\cdot)) = \frac{q}{2}\log\frac{n}{2\pi e} - \int \pi(\theta)\log\frac{\pi(\theta)}{|\Lambda^{-1}(\theta)|^{1/2}}d\theta + o(1).$$

So, the reference prior for the EL under relative entropy is

$$\pi^*_{KL}(\theta) \propto |\Lambda^{-1}(\theta)|^{1/2}.$$

We comment that the proof of the asymptotic expression is a sequence of asymptotically valid approximations whose convergence identifies the leading terms. The highest order term depending on the prior is optimized in the usual way to give the reference prior. The same comment applies to Theorems 2 and 3 below.

Proof: Recall $l_i(\theta) = \log w_i(\theta)$ and $l_i^{(2)}(\theta) = \partial^2 l_i(\theta)/(\partial\theta\partial\theta')$ and consider the limit of the mean of the $l_i^{(2)}$s. As in the proofs of Theorem 1 and 2 in Yuan et al. (2009), we have

$$\begin{aligned} w_i(\theta) &= \frac{1}{n}\frac{1}{1+t'g(x_i,\theta)} = \frac{1}{n}\Big(1 - t'g(x_i,\theta) + g'(x_i,\theta)g(x_i,\theta)O(n^{-1}(\log\log n))\Big) \\ &= \frac{1}{n}\Big(1 - B_n'g(x_i,\theta) + ||g(x_i,\theta)||o(n^{-(1-1/\alpha)}(\log\log(n)) \\ &\quad + g'(x_i,\theta)g(x_i,\theta)O(n^{-1}(\log\log n))\Big), \end{aligned}$$

where

$$B_n = \left(\frac{1}{n}\sum_{i=1}^n g(x_i,\theta)g'(x_i,\theta)\right)^{-1}\frac{1}{n}\sum_{i=1}^n g(x_i,\theta).$$

By using the Taylor expansion $\ln(1+x) \approx x$ on (4.6) we get

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n l_i(\theta) &= -\log n - B_n'\frac{1}{n}\sum_{i=1}^n g(x_i,\theta) + \frac{1}{n}\sum_{i=1}^n ||g(x_i,\theta)||o(n^{-(1-1/\alpha)}(\log\log(n)) \\ &\quad \frac{1}{n}\sum_{i=1}^n g'(x_i,\theta)g(x_i,\theta)O(n^{-1}(\log\log n)) + O(n^{-1}\log\log n). \end{aligned} \tag{4.6}$$

Taking second derivatives in (4.6) by using the product rule gives

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta) &= -\Big\{2[\frac{\partial^2}{\partial\theta\partial\theta'}\frac{1}{n}\sum_{i=1}^{n} g'(x_i,\theta)]\Big(\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)g'(x_i,\theta)\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta) \\
&\quad +2[\frac{\partial}{\partial\theta}\frac{1}{n}\sum_{i=1}^{n} g'(x_i,\theta)][\frac{\partial}{\partial\theta'}\Big(\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)g'(x_i,\theta)\Big)^{-1}]\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta) \\
&\quad +\frac{1}{n}\sum_{i=1}^{n} g'(x_i,\theta)[\frac{\partial^2}{\partial\theta\partial\theta'}\Big(\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)g'(x_i,\theta)\Big)^{-1}]\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta) \\
&\quad +[\frac{\partial}{\partial\theta}\frac{1}{n}\sum_{i=1}^{n} g'(x_i,\theta)]\Big(\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)g'(x_i,\theta)\Big)^{-1}[\frac{\partial}{\partial\theta'}\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)] \\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}||g(x_i,\theta)||o(n^{-(1-1/\alpha)}(\log\log(n)) \\
&\quad +[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}g'(x_i,\theta)g(x_i,\theta)+1]O(n^{-1}(\log\log n))\Big\}. \qquad (4.7)
\end{aligned}$$

By the strong law of large numbers, $\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta) \to Eg(X,\theta) = 0$ a.s., thus for any $\theta_n \to \theta$ (P or a.s.), only the fourth term on the right of (4.7) above is asymptotically non-zero. This gives

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n) &\to -\lim_n[\frac{\partial}{\partial\theta}\frac{1}{n}\sum_{i=1}^{n} g'(x_i,\theta)]\Big(\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)g'(x_i,\theta)\Big)^{-1}[\frac{\partial}{\partial\theta'}\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)] \\
&= -D'(\theta)\Omega^{-1}(\theta)D(\theta) = -\Lambda(\theta), \quad (P \text{ or } a.s.). \qquad (4.8)
\end{aligned}$$

We use (4.8) in the following Laplace expansion argument.

By a second order Taylor expansion in Lagrange form, we have

$$p(x^n|\theta) = \exp\{\sum_{i=1}^{n} l_i(\hat\theta_n) + \frac{1}{2}n(\hat\theta_n-\theta)'[\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)](\hat\theta_n-\theta)\}, \qquad (4.9)$$

where θ_n is betwen $\hat\theta_n$ and θ. Similarly,

$$m(x^n) = \int \pi(\theta)\exp\{\sum_{i=1}^{n} l_i(\hat\theta_n) + \frac{1}{2}n(\hat\theta_n-\theta)'[\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)](\hat\theta_n-\theta)\}d\theta. \qquad (4.10)$$

So,

$$\log\frac{p(x^n|\theta)}{m(x^n)} = \log\frac{\exp\{\frac{1}{2}n(\hat\theta_n-\theta)'[\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)](\hat\theta_n-\theta)\}}{\int\pi(\alpha)\exp\{\frac{1}{2}n(\hat\theta_n-\alpha)'[\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)](\hat\theta_n-\alpha)\}d\alpha}.$$

Let $\phi(\cdot|\hat{\theta}_n, \Lambda)$ be the q-dimensional normal density with mean $\hat{\theta}_n$ and covariance matrix Λ. Now, for any $\delta > 0$,

$$\begin{aligned}
\int & \pi(\alpha)\exp\{\frac{1}{2}n(\hat{\theta}_n-\alpha)'[\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)](\hat{\theta}_n-\alpha)\}d\alpha \\
= & \int_{||\alpha-\hat{\theta}_n||\le\delta} \pi(\alpha)\exp\{\frac{1}{2}n(\hat{\theta}_n-\alpha)'[\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)](\hat{\theta}_n-\alpha)\}d\alpha \\
& + \int_{||\alpha-\hat{\theta}_n||>\delta} \pi(\alpha)\exp\{\frac{1}{2}n(\hat{\theta}_n-\alpha)'[\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)](\hat{\theta}_n-\alpha)\}d\alpha. \qquad (4.11)
\end{aligned}$$

Write $\Lambda_n(\theta_n)^{-1} = -[\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)]$, with θ_n and $\theta_{1,n}$ in the ball $B(\hat{\theta}_n, \delta) = \{\alpha : ||\alpha - \hat{\theta}_n|| \le \delta\}$. Then, the first term on the right in (4.11) is

$$\begin{aligned}
& \pi(\theta_{1,n}) \int_{||\alpha-\hat{\theta}_n||\le\delta} \exp\{-\frac{1}{2}n(\hat{\theta}_n-\alpha)'\Lambda_n^{-1}(\theta_n)(\hat{\theta}_n-\alpha)\}d\alpha \\
= & \pi(\theta_{1,n})(2\pi)^{q/2}n^{-q/2}|\Lambda_n(\theta_n)|^{-1/2} \int_{||\alpha||\le\delta\sqrt{n}} \phi(\alpha|0, I_q)d\alpha \\
\sim & \pi(\theta)(2\pi)^{q/2}n^{-q/2}|\Lambda^{-1}(\theta)|^{1/2}, \qquad (4.12)
\end{aligned}$$

since $\delta > 0$ is arbitrary. To deal with the second term in (4.11), note that it equals

$$\pi(\theta_{2,n})(2\pi)^{q/2}n^{-q/2}|\Lambda_n(\theta_n)|^{-1/2} \int_{||\alpha||>\delta\sqrt{n}} \phi(\alpha|0, I_q)d\alpha = o(n^{-q/2}), \qquad (4.13)$$

for $\theta_{2,n}$ and $\theta_{3,n}$ in $B^c(\hat{\theta}_n, \delta)$ since $\pi(\cdot)$ and $\Lambda_n^{-1}(\cdot)$ are bounded by R6.

By Theorem 0, observe that $Y_n = n(\hat{\theta}_n - \theta)'\Lambda(\theta)(\hat{\theta}_n - \theta) \xrightarrow{D} \chi_q^2$ under $p(x^n|\theta)$. Since $E\chi_q^2 = q$ for $\epsilon > 0$, we can find $M > 0$ such that $|E[\chi_q^2 I(\chi_q^2 \le M)] - q| < \epsilon$. Weak convergence gives $E[Y_n I(Y_n \le M)] \to E[\chi_q^2 I(\chi_q^2 \le M)]$. Provided $n(\hat{\theta} - \theta)$ is uniformly integrable in P_θ, uniformly for θ, we have $E(Y_n) \to E(\chi_q^2) = q$ as $\epsilon \to 0$.

Using (4.12) and (4.13) in (4.11) and the result from Theorem 0, we have

$$\begin{aligned}
E_{m_n} D(\pi(\cdot|X^n)||\pi(\cdot)) &= \int\int \pi(\theta)p(x^n|\theta)\log\frac{p(x^n|\theta)}{m(x^n)}d\mu(x^n)d\theta \\
&\sim -\int \pi(\theta)E_{p(x^n|\theta)}\left(\frac{1}{2}n(\hat{\theta}_n-\theta)'\Lambda(\theta)(\hat{\theta}_n-\theta)\right)d\theta \\
&\quad + \frac{q}{2}\log\frac{n}{2\pi} - \int \pi(\theta)\log\pi(\theta)d\theta - \frac{1}{2}\int \pi(\theta)\log|\Lambda^{-1}(\theta)|d\theta \\
&\sim -\frac{q}{2} + \frac{q}{2}\log\frac{n}{2\pi} - \int \pi(\theta)\log\pi(\theta)d\theta - \frac{1}{2}\int \pi(\theta)\log|\Lambda^{-1}(\theta)|d\theta.\square
\end{aligned}$$

5 Hellinger Reference Prior

Next, we state and prove the analogous result for the Hellinger distance.

Theorem 2. *Assume R1–R6. Then*

$$E_{m_n}H(\pi(\cdot|x^n),\pi(\cdot)) = (2\pi/n)^{q/4}E(\exp\{\frac{1}{4}\chi^2_q\})\int|\Lambda^{-1}(\theta)|^{1/4}\pi^{3/2}(\theta)d\theta + o(\frac{1}{n^{q/4}}).$$

So, the reference prior for the EL under the Hellinger metric is

$$\pi^*_H(\theta) \propto |\Lambda(\theta)|^{1/2}.$$

Note that the reference prior under the Hellinger distance is the inverse of the reference prior under the relative entropy.

Proof: It is easy to see that

$$E_{m_n}H(\pi(\cdot|x^n),\pi(\cdot)) = 2\bigg(1-\int\int\pi(\theta)\sqrt{m(x^n)p(x^n|\theta)}d\mu(x^n)d\theta\bigg). \tag{5.14}$$

Recalling (4.9) and (4.10) from the proof of Theorem 1, we set up a slight extension of Laplace's method by expanding the prior to second order. So, let $\pi^{(1)}(\alpha) = \partial\pi(\alpha)/\partial\alpha$ and $\pi^{(2)}(\alpha) = \partial^2\pi(\alpha)/[\partial\alpha\partial\alpha']$ and recall that $\int\alpha' A\alpha\phi(\alpha|0,I_q)d\alpha = tr(A)$. Then, taking convergences in $p(x^n|\theta)$, we have

$$\begin{aligned}
&\int\pi(\alpha)\exp\{\frac{1}{2}n(\hat\theta_n-\alpha)'[\frac{1}{n}\sum_{i=1}^n l_i^{(2)}(\theta_n)](\hat\theta_n-\alpha)\}d\alpha\\
=\;& (2\pi/n)^{q/2}|\Lambda^{-1}(\theta_n)|^{1/2}\\
\times\;& \int\bigg(\pi(\hat\theta_n)+(\alpha-\hat\theta_n)'\pi^{(1)}(\hat\theta_n)+\frac{1}{2}(\alpha-\hat\theta_n)'\pi^{(2)}(\theta_{2,n})(\alpha-\hat\theta_n)\bigg)\phi(\alpha|\hat\theta_n,\Lambda^{-1}(\theta_n)/n)d\alpha\\
=\;& (2\pi/n)^{q/2}|\Lambda^{-1}(\theta_n)|^{1/2}\int\bigg(\pi(\hat\theta_n)+\frac{1}{n}\alpha'\Lambda^{-1/2'}(\theta_n)\pi^{(2)}(\theta_{2,n})\Lambda^{-1/2}(\theta_n)\alpha\bigg)\phi(\alpha|0,I_q)d\alpha\\
\sim\;& (2\pi/n)^{q/2}|\Lambda^{-1}(\theta_n)|^{1/2}\pi(\hat\theta_n)\bigg(1+\frac{1}{2n}tr[\Lambda^{-1/2'}(\theta)\frac{\pi^{(2)}(\theta)}{\pi(\theta)}\Lambda^{-1/2}(\theta)]\bigg)\\
\sim\;& (2\pi/n)^{q/2}|\Lambda^{-1}(\theta)|^{1/2}\pi(\hat\theta_n)\bigg(1+\frac{1}{2n}tr[\Lambda^{-1/2'}(\theta)\frac{\pi^{(2)}(\theta)}{\pi(\theta)}\Lambda^{-1/2}(\theta)]\bigg)\\
\sim\;& (2\pi/n)^{q/2}|\Lambda^{-1}(\theta)|^{1/2}\pi(\theta).
\end{aligned}\tag{5.15}$$

The key term in (5.14) is

$$\int\int\pi(\theta)\sqrt{m(x^n)p(x^n|\theta)}d\mu(x^n)d\theta = \int\int\sqrt{\frac{m(x^n)}{p(x^n|\theta)}}\pi(\theta)p(x^n|\theta)d\mu(x^n)d\theta.$$

So, using the square root of the ratio of (4.9) to (4.10), and (5.15) we have that

$$\begin{aligned}
&\int\int \pi(\theta)\sqrt{m(x^n)p(x^n|\theta)}d\mu(x^n)d\theta \\
\sim\ & (2\pi/n)^{q/4}\int\left(\int |\Lambda^{-1}(\theta)|^{1/4}\pi^{3/2}(\theta)\exp\{\frac{1}{4}n(\hat{\theta}_n-\theta)'\Lambda(\theta)(\hat{\theta}_n-\theta)\}p(x^n|\theta)d\mu(x^n)\right)d\theta
\end{aligned}$$

By Theorem 0, $\hat{\theta}_n \to \theta$ a.s. and $n(\hat{\theta}_n-\theta)'\Lambda(\theta_n)(\hat{\theta}_n-\theta) \xrightarrow{D} \chi^2_q$ under $p(x^n|\theta)$. So, if the convergence of the exponent to χ^2_q is uniform over θ, we get

$$\int\int \pi(\theta)\sqrt{m(x^n)p(x^n|\theta)}d\mu(x^n)d\theta \sim (2\pi/n)^{q/4}E[\exp\{\frac{1}{4}\chi^2_q\}]\int |\Lambda^{-1}(\theta)|^{1/4}\pi^{3/2}(\theta)d\theta,$$

as claimed and

$$\pi^*(\theta) = \arg\min_{\pi}\{\int |\Lambda^{-1}(\theta)|^{1/4}\pi^{3/2}(\theta)d\theta \quad \text{subject to} \quad \int \pi(\theta)d\theta = 1\}.$$

Using Lagrange multipliers and taking derivatives of $\int |\Lambda^{-1}(\theta)|^{1/4}\pi^{3/2}(\theta)d\theta - \lambda\int\pi(\theta)d\theta$ with respect to $\pi(\theta)$ for fixed θ, we get

$$\frac{3}{2}|\Lambda(\theta)|^{-1/4}\pi^{1/2}(\theta) - \lambda = 0, \quad \text{or} \quad \pi(\theta) \propto |\Lambda(\theta)|^{1/2}.\square$$

6 Chi-square Reference Prior

The following result under the Chi-square distance is analogous to Clarke and Sun (1997) and Ghosh et al. (2009), however, the solution is hard to obtain explicitly.

Theorem 3. *Assume R1–R6. Then*

$$E_{m_n}\chi^2(\pi(\cdot|x^n),\pi(\cdot)) = (\frac{n}{2\pi})^{q/2}E[\exp\{-\frac{1}{2}\chi^2_q\}]\int |\Lambda(\theta)|^{1/2}d\theta - n^{(q-2)/2}2^{(q+2)/2}\pi^{q/2}$$

$$\times E[\exp\{-\frac{1}{2}\chi^2_q\}]\int |\Lambda(\theta)|^{1/2}tr[\Lambda^{-1/2'}(\theta)\frac{\pi^{(2)}(\theta)}{\pi(\theta)}\Lambda^{-1/2}(\theta)]d\theta + o(n^{(q-2)/2}).$$

So, the reference prior for the EL under the Chi-square distance is

$$\pi(\cdot) = \arg\min_{\pi(\cdot)}\int |\Lambda(\theta)|^{1/2}tr[\Lambda^{-1/2'}(\theta)\frac{\pi^{(2)}(\theta)}{\pi(\theta)}\Lambda^{-1/2}(\theta)]d\theta, \quad \textit{subject to} \quad \int\pi(\theta)d\theta = 1.$$

Proof: As in Theorem 2, let $\pi^{(1)}(\theta) = \partial\pi(\theta)/\partial\theta$ and $\pi^{(2)}(\theta) = \partial^2\pi(\theta)/[\partial\theta\partial\theta']$ and recall that $\int\theta'A\theta\phi(\theta|0,I_q)d\theta = tr(A)$. Now, when $p(x^n|\theta)$ defines the mode of convergence,

Taylor expanding gives

$$
\begin{aligned}
\int \quad & \pi(\alpha)\exp\{\frac{1}{2}n(\hat{\theta}_n-\alpha)'[\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)](\hat{\theta}_n-\alpha)\}d\alpha \\
\sim \quad & (2\pi/n)^{q/2}|\Lambda^{-1}(\theta)|^{1/2} \\
\times \quad & \int\left(\pi(\hat{\theta}_n)+(\alpha-\hat{\theta}_n)'\pi^{(1)}(\hat{\theta}_n)+\frac{1}{2}(\alpha-\hat{\theta}_n)'\pi^{(2)}(\theta_{2,n})(\alpha-\hat{\theta}_n)\right)\phi(\alpha|\hat{\theta}_n,\Lambda^{-1}(\theta)/n)d\alpha \\
= \quad & (2\pi/n)^{q/2}|\Lambda^{-1}(\theta)|^{1/2} \\
\times \quad & \int\left(\pi(\hat{\theta}_n)+\frac{1}{\sqrt{n}}\alpha'\Lambda^{-1/2'}(\theta)\pi^{(1)}(\hat{\theta}_n)+\frac{1}{2n}\alpha'\Lambda^{-1/2'}(\theta)\pi^{(2)}(\theta_{2,n})\Lambda^{-1/2}(\theta)\alpha\right)\phi(\alpha|0,I_q)d\alpha \\
\sim \quad & (2\pi/n)^{q/2}|\Lambda^{-1}(\theta)|^{1/2}\pi(\theta)\left(1+\frac{1}{2n}tr[\Lambda^{-1/2'}(\theta)\frac{\pi^{(2)}(\theta)}{\pi(\theta)}\Lambda^{-1/2}(\theta)]\right). \qquad (6.16)
\end{aligned}
$$

Using the inverse of (6.16), we have that $E_{m_n}\chi^2(\pi(\cdot|x^n),\pi(\cdot))$ equals

$$
\begin{aligned}
& \int\int \pi(\theta)\frac{p^2(x^n|\theta)}{m(x^n)}d\mu(x^n)d\theta-1 \\
= \quad & \int\int \pi(\theta)p(x^n|\theta)\frac{\exp\{\frac{1}{2}n(\tilde{\theta}_n-\theta)'[\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)](\tilde{\theta}_n-\theta)\}}{\int\pi(\alpha)\exp\{\frac{1}{2}n(\tilde{\theta}_n-\alpha)'[\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)](\tilde{\theta}_n-\alpha)\}d\alpha}d\mu(x^n)d\theta-1 \\
\sim \quad & (\frac{n}{2\pi})^{q/2}\int\int|\Lambda(\theta)|^{1/2}\left(1-\frac{1}{2n}tr[\Lambda^{-1/2'}(\theta)\frac{\pi^{(2)}(\theta)}{\pi(\theta)}\Lambda^{-1/2}(\theta)]\right) \\
\times \quad & \exp\{\frac{1}{2}n(\tilde{\theta}_n-\theta)'[\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\tilde{\theta}_n)](\tilde{\theta}_n-\theta)\}p(x^n|\theta)d\mu(x^n)d\theta-1 \\
\sim \quad & (\frac{n}{2\pi})^{q/2}E[\exp\{-\frac{1}{2}\chi_q^2\}]\int|\Lambda(\theta)|^{1/2}\left(1-\frac{1}{2n}tr[\Lambda^{-1/2'}(\theta)\frac{\pi^{(2)}(\theta)}{\pi(\theta)}\Lambda^{-1/2}(\theta)]\right)d\theta.\square
\end{aligned}
$$

7 Discussion

It is seen that the reference prior for ELs under Hellinger is based on the reciprocal of the reference prior under relative entropy and that these differ from the reference prior under χ^2 which is hard to obtain explicitly. This is somewhat different from the treatment given in Ghosh et al. (2009) who obtained the Jeffreys prior for all members of the power divergence family except the Chi-square distance. Here, it is only in the relative entropy case that the reference prior is based on the transformation that makes an efficient CAN estimator converge to $N(0,I_q)$. Nevertheless, the role of Jeffreys prior is roughly analogous to $\Lambda^{-1}(\theta)=(D'(\theta)\Omega^{-1}(\theta)D(\theta))^{-1}$.

An examination of the proof of all three theorems reveals a common structure: Approximate the ratio $p(x^n|\theta)/m(x^n)$ by a Laplace's method argument, take a function of the

density ratio, and examine its limiting expectation using standard results and assumptions. Consequently, we conjecture that our basic technique extends to any Csiszar f-divergence, see Csiszar (1967), defined as $D_f(p||q) = E_p f(p/q)$ for some convex f where p and q are densities. The power divergence family (whose members often play a role in goodness-of-fit testing) is contained in this class. Many reference priors could be generated this way, however, outside the relative entropy case, reference priors do not correspond to least favorable priors.

References

Berger, J. and J. Bernardo (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc. 84*, 200–207.

Berger, J. and J. Bernardo (1991). Reference priors in a variance components problem. In P. Goel and N. Iyengar (Eds.), *Bayesian Inference in Statistics and Econometrics*. New York: Springer.

Berger, J. and J. Bernardo (1992a). On the development of reference priors. In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics IV*. Oxford: Clarendon Press.

Berger, J. and J. Bernardo (1992b). Ordered group reference priors with application to the multinomial. *Biometrika 25*, 25–37.

Berger, J., J. Bernardo, and M. Mendoza (1991). On priors that maximize expected information. In J. Klein and J. Lee (Eds.), *Recent Developments in Statistics and their Applications*. Seoul: Freedom Academy.

Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *J. Roy. Statist. Soc. B 41*, 113–147.

Clarke, B. and D. Sun (1997). Reference priors under the chi-square distance. *Sankhya 59*, 215–231.

Clarke, B. and A. Yuan (2004). Partial information reference priors: derivation and interpretations. *J. Stat. Planning Inference 123*(2), 313–345.

Csiszar, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar. 2*, 229–318.

Ghosh, J. and R. Mukerjee (1992). Noninformative priors. In e. a. Bernardo, J. (Ed.), *Bayesian Statistics IV*, Oxford, pp. 195–210. Clarendon Press.

Ghosh, M. (2009). Objective priors: A selective review. Technical report, Dept. of Statistics, Univ, of Florida.

Ghosh, M., V. Mergel, and R. Liu (2009). A generali divergence criterion for prior selection. *To apear: Ann. Inst. Stat. Math.*.

Grendar, M. and G. Judge (2009). Asymptotic equivalence of empirical likelihood and bayesian map. *Ann. Statist. 37*, 2445–2457.

Herve, A. (2007). Distance. In N. Salkind (Ed.), *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage.

Ibragimov, I. and R. Hasminsky (1973). On the information in a sample about a parameter. In *Proc. 2nd Internat. Symp. on Information Theory*, Budapest, pp. 295–309. Akademiai, Kiado.

Lazar, N. (2003). Bayesian empirical likelihood. *Biometrika 90*, 319–326.

Liu, R. and M. Ghosh (2009). Objective priors: A selective review. Technical report, Dept. of Statistics, Univ, of Florida.

Moon, H. and F. Schorfheide (2004). Bayesian inference for econometric models using empirical likelihood functions. In *2004 North American Winter Meetings*. Paper #284, see http://repec.org/esNAWM04/up.25738.1049064209.pdf: Econometric Society.

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

Owen, A. (1990). Empirical likelihood for confidence regions. *Ann. Statist. 18*, 90–120.

Owen, A. (1991). Empirical likelihood for linear models. *Ann. Statist. 19*, 1725–1747.

Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *Ann. Statist. 22*, 300–325.

Shannon, C. (1948a). A mathematical theory of communication, part i. *Bell Syst. Tech. J. 27*, 379423.

Shannon, C. (1948b). A mathematical theory of communication, part ii. *Bell Syst. Tech. J 27*, 623656.

Sun, D. and J. Berger (1998). Reference priors with partial information. *Biometrika 85*, 55–71.

Yuan, A., G. Zheng, and J. Xu (2009). On empirical likelihood statistical functions. Technical report, Human Genome Project, Howard University.

Zhang, Z. (1994). *Discrete Noninformative Priors*. Ph. D. thesis, Department of Statistics, Yale University.