Bayesian Properties of Minimum Divergence and Generalized Empirical Likelihood Methods (PRELIMINARY DRAFT)*

Giuseppe Ragusa
Rutgers University †
gragusa@rci.rutgers.edu

March 22, 2006

Abstract

When the object of the statistical analysis is the estimation of an economic model, the choice of a likelihood function should be coherent with the economic model under investigation. Many econometric models provides the researcher with weak "structural prediction" about the parameter of interest and the data. Econometric models specified through moment conditions and usually estimated by Generalized Method of Moments (GMM) belong to this class. This paper show that valid likelihoods (in a Bayesian sense) can be constructed by substituting the parametric likelihood in the Bayes theorem with some (empirical) likelihoods whose functional forms have a close relationship with the weights generated by Minimum Divergence (MD) and Generalized Empirical Likelihood (GEL). These likelihoods are obtained by eliciting a prior distribution for the nuisance parameters that is maximally uninformative but contains information about the moment conditions. Integrating over the nuisance parameters with respect to this distribution gives likelihoods that have a relationship with the weights that MD/GEL methods assign to a given observation. The likelihoods obtained are proper only when all the possible outcomes of the underlying random vector is observed. This issue is investigated by applying a simply methodology proposed by Monahan and Boos (1992). Higher order properties of the maximum posterior estimators are also discussed.

Keywords: Bayesian Likelihood, GMM, GEL, Minimum Divergence, Higher Order asymptotic.

*I wish to thank Graham Elliott, Hal White, Ulrich Muller, Dale Poirer, Francesca Mazzolari for helpful discussions on an earlier draft and seminar participants at Princeton University, Columbia University (Greater New York Metropolitan Area Econometrics Colloquium) and Brown.

†Department of Economics, New Jersey Hall, New Brunswick, NJ 08901.

1 Introduction and Motivations

The fundamental component of statistical analysis in a Bayesian framework is the distribution of the data to be observed given the parameter, the likelihood function. Choice of a likelihood function amounts to choice of a family of probability distributions, one for each element of the parameter space. When the object of the statistical analysis is the estimation of an economic model, the choice should be coherent with the economic model under investigation. Most economic theory provides the researcher with only weak "structural predictions" about the parameter of interest and the data. In such a situation it seems inappropriate to specify a likelihood that would involve beliefs that are not intrinsic to the model. While it is customary to achieve a formulation of a likelihood through a fully parametric specification, many econometric models do not lend themselves to a parametric specification. For instance, econometric models specified through moment conditions are difficult to parameterize unless further structure is imposed on the model. Even when it is possible to specify a parametric likelihood, robustness concerns may discourage its use.

When the only information takes the form of a set of moment conditions, classical statistical analysis is carried out by the well known Generalized Method of Moments (GMM) framework of Hansen (1982). The asymptotic properties of GMM are well understood and well documented; see, for example, Newey and McFadden (1994). Recently, many important papers have studied the behavior of estimators and tests statistics based on GMM when the number of instruments is large and/or the moment condition identifies the parameters of interest only weakly; see, for example, Stock and Wright (2000), Kleibergen (2005) and Han and Phillips (2006). On the other hand, there is little work on Bayesian estimation in a moment conditions setting. This is particularly puzzling since, as pointed out by Sims (1996), GMM procedures are often used for real decision making and it is inevitable that classical confidence bands are going to be interpreted as if they were the posterior probability credible regions.

There has been some attempts at giving a Bayesian interpretation to GMM. Kim (2002) attempts at justifying the use of the GMM objective function as a limited information likelihood through an asymptotic argument based on the minimization of the Kullback-Leibler distance. Chamberlain and Imbens (2003) propose a "multinomial approach" to inference for moment conditions models. This approach is not based on GMM, rather it is reminiscent of the Bayesian bootstrap of Rubin (1981). Although it has a number of attractive features, this method appears to be lacking some of the properties that pertain to proper Bayesian inference.

Lazar (2003) attempts to give a Bayesian interpretation to the empirical likelihood framework developed by Owen (1988) and adapted to estimation of moment conditions models by Imbens (1997) and Qin and Lawless (1994). Lazar considers the "natural" pro-

cedure of using as likelihood in the posterior distribution the profiled empirical likelihood (EL). The process consists in multiplying the empirical likelihood by a prior distribution and then renormalizing in order to obtain a proper density function. As pointed out by Schennach (2005), the use of the profiled empirical likelihood lacks a sound probabilistic interpretation. By putting a prior on the space of distributions that support the moment conditions, Schennach (2005) obtains an integrated likelihood, the so called "Bayesian Exponentially-Tilted Empirical Likelihood" (BETEL). The BETEL does not coincide with the profiled EL, rather it combines features of the profiled EL with features of the profiled Exponential Tilted (ET) distribution of Efron (1981), Kitamura (1997) and Imbens (1997).

This paper aims at extending the previous results and at providing a probabilistic justification for the use of generalized "empirical" likelihood. The form of these likelihoods has interesting connections with the Minimum Divergence methods studied in Ragusa (2005) and their GEL counterpart analyzed in Newey and Smith (2004). The procedure through which the likelihoods are obtained puts a prior on the whole simplex of multinomial weights. An initial prior is transformed so to have the maximum entropy among the priors that satisfy the moment conditions. The nuisance parameters are then integrated out with respect to this prior. The relationship between the MD/GEL weights and the likelihood is indexed by the initial choice of the prior.

Similarly to the Bayesian Bootstrap of Rubin and the adaptation to moment restrictions models by Chamberlain and Imbens (2003), the procedure is fully Bayesian, and hence valid, when all the distinct values of the underlying random variable have been observed. When this is not the case, the procedure is not formally Bayesian and the issue of the Bayesian validity of the derived likelihood functions must be addressed. Using a simple approach described in Monahan and Boos (1992) that relies on the concept of correct coverage of the posterior distributions, we provide evidence that some of these likelihoods are indeed valid for Bayesian inference.

The plan of the paper is as follows. Section 2 describes the approaches taken by previous literature to embed models specified through moment conditions in a Bayesian likelihood framework. Section 3 describes the proposed new approach and its properties. Section 4 discusses the validity of the likelihoods obtained for Bayesian inference. Section 5 digs into the asymptotic properties of the Bayesian estimators and inference based on the posterior. Section 6 contains a simple application of the methods developed in Section 3. Finally, Section 6 concludes.

2 Limited Information Likelihoods

In many economic applications interest centers on parameters that have an economic interpretation, but the full model may involve a nuisance component about which there is

no available *apriori* knowledge. This situation arises in the context of models specified through moment conditions. Given a parameter vector $\theta \in \Theta$, where Θ is a compact subset of $\mathbb{R}^k$, and given a function $q(\cdot, \theta) : \mathbb{R}^k \to \mathbb{R}^m$, $m \geq k$, a moment conditions model is an econometric model defined by

$$\int q(W, \theta_0) \mathrm{d}F_0 = 0 \tag{1}$$

where W is a random variable with distribution F_0. The model is complemented with an identification assumption: for every $\theta \neq \theta_o \in \Theta$, $\| \int q(W, \theta) \mathrm{d}F_o \| > 0$. In the standard GMM framework, a point estimate of θ_0 is obtained by solving the following optimization problem

$$\min_{\theta \in \Theta} n q_n(\theta)' \mathcal{W}_n(\theta) q_n(\theta) \tag{2}$$

where $q_n(\theta) = \sum_{i=1}^n q(w_i, \theta)/n$ and $\mathcal{W}_n(\theta) = \mathcal{W}(\theta) + o_p(1)$ uniformly in $\theta \in \Theta$ and $\mathcal{W}(\theta)$ is positive definite uniformly in $\theta \in \Theta$. Under standard (e.g. Hansen (1982)) and non standard (e.g. Pakes and Pollard (1989)) regularity conditions, the solution to the above optimization step, say $\hat{\theta}_n$, is a $\sqrt{n}$-consistent estimator of θ_0, i.e. $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1)$ with asymptotically normal distribution. Setting $\mathcal{W}_n(\theta_o) = V + o_p(1)$, with $V = E[q(W, \theta_0) q(W, \theta_0)']$, implies the generalized information equality.

The usual formula for Bayesian inference is

$$p(\theta | w_1, \ldots, w_n) \propto p(\theta) \ell(w_1, \ldots, w_n | \theta)$$

where $p(\theta)$ is the prior density for the parameter θ, $\ell(w_1, \ldots, w_n | \theta)$ is the likelihood for the observed *datum* $(w_1, w_2, \ldots, w_n)$ and $p(\theta | w_1, \ldots, w_n)$ is the posterior distribution of θ. In a fully parameterized model, the likelihood is the conditional density $f(w_1, \ldots, w_n | \theta)$ of the random variable W given θ. When dealing with models as the one specified in (1), the full likelihood is usually not available.

Kim (2002) discusses how the GMM can be embedded in a likelihood-based Bayesian inference framework. The information contained in the moment restriction is equivalent to the information contained in

$$\lim_{n \to \infty} E_{F_0} \left[n q_n(\theta)' V^{-1} q_n(\theta) \right] = m$$

The idea is to work, for each $\theta \in \Theta$, with the set of probability measures that are absolutely continuous with respect to F_0 and consistent with the above limiting restriction. This is equivalent to consider the set

$$\mathscr{F}(\theta) = \left\{ Q : \lim_{n \to \infty} E_Q [n q_n(\theta)' V^{-1} q_n(\theta)] = m \right\}$$

The limited information likelihood based on GMM is the finite sample counterpart of the measure $G(\theta)^* \in \mathscr{F}(\theta)$ that minimizes the Kullback-Leibler distance from F_0, that is

$$G(\theta)^* = \arg \min_{G \in \mathscr{F}(\theta)} \int \log(\mathrm{d}G/\mathrm{d}F_0)\mathrm{d}G$$

The sample counterpart of the density $g^*(t, w^n) = \mathrm{d}G/\mathrm{d}F_o$ that solves the Kullback-Leibler problem is given by

$$g_n^*(\theta, w_1, \ldots, w_n) \propto \exp\left\{-\frac{1}{2} n q_n(\theta) \hat{V}_n^{-1} q_n(\theta)\right\}$$

Kim suggests using $g_n^*(\theta)$ as the likelihood in the Bayes theorem to obtain the quasi-posterior

$$p^*(\theta|w_1, \ldots, w_n) \propto g^*(\theta, w_1, \ldots, w_n) p(\theta) \tag{3}$$

A quasi-posterior can be easily obtained from models that are specified through a large number of moment conditions or through non-smooth moment conditions. Unfortunately, the procedure is only justified asymptotically and it is not clear whether for a given sample the quasi-posterior in (3) delivers valid Bayesian inference.

The quasi-posterior (3) has also been considered by Chernozhukov and Hong (2003) in their study of Laplace type estimators. They focus on generating a class of estimators that are defined as statistics of the quasi-posterior distribution in (3) and have desirable asymptotic classical properties, such as consistency, asymptotic normality and asymptotic efficiency. The Laplace-type of estimators are implemented by simulating from (3) through Markov Chain Monte Carlo methods. The approach of Kim (2002) is very appealing from a computationally point of view.

Chamberlain and Imbens (2003) take a different approach. They consider sampling from a posterior distribution obtained by assuming a Dirichlet prior on the parameters of the multinomial distribution. If $\beta = (\beta_1, \beta_2, \ldots, \beta_J)$ denote the vector of multinomial probability, a Dirichlet prior on β consists of

$$p(\beta) = \prod_{j=1}^{J} \beta_j^{b_j - 1}$$

Given this prior and by conjugacy of the Dirichlet process with the multinomial distribution, the posterior is proportional to

$$\prod_{j=1}^{J} \beta^{n_j + b_j - 1}$$

When $b_j \to 0$ for any j, sampling from the above posterior under the moment constraints

amounts to solve

$$\sum_{i}^{n} q(w_i, \theta^{(l)}) v_i^{(l)} = 0$$

for $\{v_i^{(l)}\}_i^n$ i.i.d. exponential random variables. Repeating this procedure for $l = 1, \ldots, L$ gives L independent draws from the posterior distribution of θ. The are two main problems with this approach. First, it is not clear how to incorporate prior knowledge about θ. A prior on θ is elicited implicitly by the choice of Dirichlet distribution on the multinomial weights and it is not easy to ascertain whether the specific prior on β is adding unrequested information about θ. To measure the informativeness of the Dirichlet distribution, Chamberlain and Imbens (2003) propose calculating the (expected) posterior distribution given a small number of observations and comparing it with the full posterior distribution. The second problem is that since drawing from the posterior distribution entails solving the moment equations, the method runs into difficulties when dealing with over-identified models, that is when the dimension of $q(\cdot, \cdot)$ is higher than the dimension of θ. In this case the posterior can be simulated by solving, for each set of $\{v_i^{(l)}\}_i^n$ i.i.d. exponential random variables, the weighted sample equation $\sum_{i=1}^n \tilde{q}(w_1, \delta) v^{(l)} = 0$, where $\delta = (\theta_0, \theta_1, \Gamma_0, \Gamma_1, \Delta)$ and $\tilde{q}(w_i, \theta)$ is the augmented moment function defined as

$$\tilde{q}(w, \delta) = \begin{pmatrix} \Gamma_0' q(w, \theta_0) \\ \text{vec}(\partial q(w, \theta_0)/\partial \theta' - \Gamma_0) \\ vec(q(w, \theta_0) q(w, \theta_0)' - \Delta) \\ \Gamma_1' \Delta^{-1} q(w, \theta_1) \\ \text{vec}(\partial q(w, \theta_1)/\partial \theta' - \Gamma_1) \end{pmatrix}$$

Ignoring the problems arising from the existence of multiple roots –problem that may arise in a just identified nonlinear setting– the dimension of $\tilde{q}(w, \theta)$ and the method may nevertheless become very computationally demanding even for models with a relatively small number of parameters and/or of moment restrictions.

Lazar (2003) proposes to replace the likelihood in the formula for the posterior with the profiled empirical likelihood, that is

$$\ell(w^n|\theta) = \left\{ \max_{\{\pi_i\}_i^n} \prod_{i=1}^{n} \pi_i, \text{ subject to } \sum_{i}^{n} \pi_i q(w_i, \theta) = 0, \sum_{i}^{n} \pi_i = 1 \right\}$$

Although profiled likelihoods have been used to approximate marginal posteriors, there is no reason to expect that a profile likelihood will behave like a marginal posterior. For example, a theoretical obstacle is that the profiled empirical likelihood deals with the nuisance parameters $\{\pi\}_{i=1}^n$ by maximizing over them, rather of integrating them out with respect to some reference prior distribution.

3 GEL/MD Bayesian Posterior

In this section we offer a well defined probabilistic justification for using, in the posterior, likelihoods that are obtained by multiplying the weights of the Minimum Divergence problem, as studied by Ragusa (2005). Let $\gamma(\cdot)$ be a function satisfying, for any $a_\gamma < 1 < b_\gamma$ the following conditions

$$\begin{aligned} &\gamma : (a_\gamma, b_\gamma) \to \mathbb{R};\ \gamma(1) = 0;\ \frac{\partial \gamma(x)}{\partial x}\,|_{x=1} = 0; \\ &\frac{\partial^2 \gamma(x)}{\partial x} > 0,\ x \in (a_\gamma, b_\gamma);\ \frac{\partial^2 \gamma(x)}{\partial^2 x}\,|_{x=1} = 1 \end{aligned} \tag{A.1}$$

For a given θ, the minimum divergence problem is formally defined as

$$\min_{\{\pi_i\}_{i=1}^n} \left\{ \frac{1}{n} \sum_{i=1}^n \gamma(n\pi_i),\ \text{subject to } \sum_{i=1}^n \pi_i q(w_i, \theta) = 0;\ \sum_{i=1}^n \pi_i = 1 \right\} \tag{4}$$

For $\gamma(x) = -\log(x) + x - 1$, $\gamma(x) = \exp(x) - x - 1$ and $\gamma(x) = x^2/2$ the minimum divergence problem corresponds to the Empirical Likelihood (EL), the Exponential Tilting (ET) and the Continuous Updating (CU), respectively. Let $\pi^\gamma(\theta) = (\pi_1^\gamma(\theta), \ldots, \pi_n^\gamma(\theta))$ denote the solutions of the MD problem for a given θ and for a given function $\gamma(\cdot)$. The weights obtained as solution to (4) are related to the Generalized Empirical Likelihood of Newey and Smith (2005). For a given θ, the GEL problem is defined as

$$\min_{\lambda \in \Lambda(\theta)} \frac{1}{n} \sum \psi(\lambda' q(w_i, \theta)) \tag{5}$$

As showed by Ragusa (2005), there is a precise relationship between (4) and (5). Let $\gamma_r(x) = \partial^r \gamma(x)/\partial^r x$, $\psi_r(y) = \partial \psi(y)/\partial y$, $y = \tilde{\gamma}_1(x)$ and $x = \tilde{\psi}(y)$ such that $\gamma_1(y) = x$ and $\psi(x) = y$. Then, for a given function $\gamma(\cdot)$ satisfying assumption **??**, $\pi_i^\gamma(\theta) = \pi_i^\psi(\theta)$ where $\pi_i^\psi(\theta) \propto \psi_1(\hat{\lambda}' q(w_i, \theta))$, $\psi(x) = x\tilde{\gamma}_1(x) - \gamma(\tilde{\gamma}_1(x))$ and $\hat{\lambda}$ is the minimand of (5). Similarly, let $\psi(\cdot)$ be a function satisfying for any $\alpha_\psi < 0 < b_\psi$ the followings

$$\begin{aligned} &\psi : (a_\psi, b_\psi) \to \mathbb{R}, \psi(0) = 0, \gamma(1) = 0;\ \frac{\partial^2 \psi(y)}{\partial^2 y} > 0, y \in (a_\gamma, b_\gamma); \\ &\frac{\partial \psi(y)}{\partial y}\,|_{y=1} = 1;\ \frac{\partial^2 \psi(y)}{\partial^2 y}\,|_{y=1} = 1 \end{aligned} \tag{A.1'}$$

Then $\pi_i^\psi(\theta) = \pi_i^\gamma(\theta)$ were $\pi^\psi(\theta) \propto \psi(\hat{\lambda}' q(w_i, \theta))$, $\pi^\gamma(\theta) \propto \tilde{\gamma}_1(\hat{\lambda}' q(w_i, \theta))$ and $\gamma(x) = x\tilde{\psi}_1(x) - \psi(\tilde{\psi}_1(x))$. A natural question is whether likelihoods proportional to $\prod_i^n \pi_i^\gamma(\theta)$, or equivalently, proportional to $\prod_{i=1}^n \pi^\psi(\theta)$, can be obtained by a formal Bayesian argument.

In a formal Bayesian framework, nuisance parameters are dealt with by integration.

Given the *datum* d, the integrated posterior is given by the following expression

$$p(\theta|d) = \int \left\{ \prod_{i=1}^{n} p(w_i|s,\theta) \right\} dP(s|\theta)\ p(\theta) \tag{6}$$

where the joint prior on (ϕ, θ), $p(\phi, \theta)$ is factorized as $dP(\phi|\theta)p(\theta)$.

We consider a probabilistic framework that is very similar to that of Chamberlain and Imbens (2003). The observations $d = (w_1, \ldots, w_n)$ are realizations from W, i.i.d. distributed according to the the family of distributions $\{F_\beta : \beta \in \mathcal{B}\}$ for some unknown value of $\beta \in \mathcal{B}$. We assume that the distributions F_β have a common finite support, $P(W = a_j) = \beta_j$, $(j = 1, \ldots, J)$, where β_j is the $j - th$ element of β and $\mathcal{B}$ is the unit simplex in $\mathbb{R}^k$. The moment condition is then given by

$$\sum_{i=1}^{n} q(a_j, \theta)\beta_j = 0$$

Let $n_j = \sum_{i=1}^{n} 1(w_i = a_j)$ be the number of sample observations equal to a_j. The likelihood of d can then be written as

$$p(d|\beta) = \prod_{j=1}^{J} \beta_j^{n_j}$$

Instead of considering the likelihood above, we consider a likelihood based on the probability given to the observations by apportioning the probabilities β_j into specific weights $\phi_i > 0$, $i = 1, \ldots, n$. Let the weights ϕ_i be implicitly defined by $\beta_j = \sum_{i=1}^{n} 1(w_i = a_j)\phi_i$. The likelihood can be rewritten in terms of ϕ_i's $p(d|\phi) = \prod_{j=1}^{n} \phi_i$ that results to be equal up to a multiplicative constant to $p(d, \beta)$. The value of likelihood above is not unique because it depends on the specific value assigned to the $\phi_i's$ consistent with $\beta_j = \sum_{i=1}^{n} 1(w_i = a_j)\phi_i$. Conveniently, we set $\phi_i = \beta_{j(i)}/n_{j(i)}$, where $\beta_{j(i)}$ is identified by $(w_i = a_j)$ and $n_{(i)} = \sum_{j=1}^{J} 1(w_i = a_j)$. Using this setting, the likelihood becomes

$$p(d|\phi) = \prod_{j=1}^{J} \beta^{n_j} = \prod_{j=1}^{J} n_j^{-n_j} \prod_{i=1}^{n} \phi_i$$

and hence

$$\prod_{i=1}^{n} \phi_i \propto \prod_{j=1}^{J} \beta^{n_j}$$

Under the assumption that all possible distinct values of W have been observed, the

population moment conditions can be rewritten in terms of the observations,

$$\begin{aligned}\sum_{i=1}^{n}\phi_i q(w_i,\theta) &= \sum_{j=1}^{J}\sum_{i=1}^{n} 1(w_i = a_j)\phi_i q(a_j,\theta) \\ &= \sum_{j=1}^{J}\beta_j q(a_j,\theta)\end{aligned}$$

The assumption that all possible distinct values of W have been observed is of course highly questionable. The same assumption is made by the Bayesian Bootstrap approach when the prior on β is set to be the improper prior proportional to $\prod \beta^{-1}$. Rubin (1981) argues that this assumption is not innocuous. Its failure may be serious and may invalidate the formal validity of the likelihoods and, consequently, their practical performance. However, an advantage of the following approach over the Bayesian Bootstrap is that the coverage validity of the resulting likelihood can be investigated.

We consider putting a prior on ϕ. The idea is to find a prior $p(\phi|\theta)$ such that the moment condition is satisfied in expectations, that is

$$E_{P(\phi|\theta)}\left[\sum_{i=1}^{n}\phi_i q(w_i,\theta)\right] = 0.$$

The prior can either be a discrete or a continuous density. Consider the following set of measures

$$\mathscr{F}_\phi(\theta) = \left\{P(\phi|\theta) : \int\left[\sum_{i=1}^{n}\eta_i q(w_i,\theta)\right] dP(\eta|\theta) = 0\right\}$$

Among the elements of $\mathscr{F}_\phi(\theta)$, we consider as prior for the nuisance parameter the distribution that has the largest entropy with respect to an initial prior distribution, say $p(x) = \prod_{i=1}^{n} p(x_i)$. Formally, we seek a solution to the following optimization problem

$$P(\phi|\theta) = \arg\max_{P(\eta|\theta)\in\mathscr{F}_\phi(\theta)}\left\{-\int\log\left(\frac{dP(\eta|\theta)}{dP(\phi)}\right) dP(\eta|\theta)\right\} \tag{7}$$

Intuitively, the prior specified by the optimization in (7), should result in a prior that incorporates the available information about β_j, expressed here in terms of $\sum_{i=1}^{n}\phi_i q(w,\theta) = 0$, but otherwise is as non-informative as possible with respect to the initial prior $p(\phi)$. The approach of entropy of the distribution supported by the moment restriction was proposed by Jaynes (1983) and Rosenkrantz(1977). Let $q_i(\theta) = q(w_i,\theta)$.

Theorem 1. *The solution to the problem* (7) *is given by*

$$P(\phi|\theta) = \frac{\exp\left(\sum_{i=1}^{n}\lambda' q_i(\theta)\phi_i\right) P(\phi)}{\int\exp\left(\sum_{i=1}^{n}\lambda' q_i(\theta)\phi_i\right) \mathrm{d}P(\phi)}$$

where λ is the solution to

$$\int \exp\left\{\sum_{i=1}^{n} \lambda' q_i(\theta)\phi_i\right\} \sum_{i=1}^{n} q_i(\theta)\phi_i \mathrm{d}P(\phi) = 0$$

The measure $p(\phi|\theta)$ needs not exist. A sufficient condition is that the λ solving the constraint in Theorem 1 belongs to the following set (see Csiszar(1975))

$$\mathscr{T}(\theta) = \left\{ v : \int \exp\left[v' \sum_{i}^{n} q(w_i, \theta)\phi_i \right] \mathrm{d}P(\phi) < \infty \right\}$$

When $\lambda \notin \mathscr{T}(\theta)$, we set $P(\phi|\theta) = 0$. Under $P(\phi|\theta)$ the expected value of the sample moment condition is zero. It also holds that $(\phi_1, \ldots, \phi_n)$ are independent as the following theorem shows.

Theorem 2. *Under $P(\phi|\theta)$ of Theorem 1, $(\phi_1, \ldots, \phi_n)$ are independent, each with distribution*

$$P(\phi_i|\theta) = \exp[\tau_i\phi_i - \varphi(\tau_i)]P(\phi)$$

where $\tau_i = \sum_{i=1}^{n} \lambda' q_i(\theta)$ and

$$\varphi(\tau) = \log \int \exp(\tau\phi)\mathrm{d}P(\phi)$$

and λ solves

$$\sum_{i=1}^{n} \int \exp(\tau_i\phi_i) q(w_i, \theta)\phi_i \mathrm{d}P(\phi_i) = 0$$

The function $\varphi(\cdot)$ in Theorem 2 is the logarithm of the moment generating function of ϕ. A property of $\varphi(\cdot)$ is that its derivative is given by

$$\varphi_1(\tau) = \int \phi[\exp(\tau\phi) - \varphi(\tau)]\mathrm{d}P(\phi) \tag{8}$$

Notice that the prior in Theorem 2 is consistent with the definition of the weights parameter ϕ. Recall that each $\phi_i = \beta_{j(i)}/n_{(i)}$ and that we would expect a prior that assigns equal probability to the events $\{W = w_i\}$ and $\{W = w_r\}$, whenever $w_i = w_r$. The maximum entropy prior results in $p(\phi_j|\theta) = p(\phi_r|\theta)$ whenever $\phi_r = \phi_j$, that is whenever two equal observations are present in the sample.

Using the independence of the prior established in Theorem 2, consider integrating out the nuisance parameter with the maximum entropy prior with respect to the reference prior

$$p(\theta|d) \;\; = \;\; \int \left\{ \prod_{i}^{n} \phi_i \right\} \prod_{i=1}^{n} \{\exp[\tau_i \phi_i - \varphi(\tau_i)]\} \, \mathrm{d}P(\phi|\theta) \; p(\theta)$$

Using the properties of the product, independence of ϕ under the maximum entropy prior and rearranging, the posterior can be written as

$$p(\theta|d) \propto \prod_{i=1}^{n} \int \phi_i \exp[\tau_i \phi_i - \varphi(\tau_i)] dP(\phi|\theta) \; p(\theta) \tag{9}$$

The expression for the posterior in (9) is very convenient, because it reduces the calculation of the posterior from a very complex expression to a formula that involves the reference prior $p(\phi)$, through the logarithm of the moment generating function as given in (8). Noticing that the terms involving the integral of the numerator of the maximum entropy prior and the nuisance parameter can be expressed as

$$\varphi_1(\tau_i) = \int \phi_i \exp[\tau_i \phi_i - \varphi(\tau_i)] \mathrm{d}P(\phi),$$

the posterior can be conveniently rewritten as

$$p(\theta|d) \propto \prod_{i=1}^{n} \varphi_1(\lambda' q_i(\theta)) \tag{10}$$

The parameter λ is implicitly defined as solution to the equation given in Theorem 2. Consider the following optimization problem

$$\min_{\lambda \in \Lambda(\theta)} \sum_{i=1}^{n} \varphi(\lambda' q(w_i, \theta)) \tag{11}$$

It is easy to verify that when $p(\phi)$ is non-degenerate, $\varphi(x)$ is strictly convex. It follows that the minimum is attained for λ solving

$$\sum_{i=1}^{n} \varphi_1(\lambda' q(w_i, \theta)) q(w_i, \theta) = 0 \tag{12}$$

Hence, the parameter λ can be equivalently defined as minimand of the optimization problem in (11) or as solution of equation (12). The set of $\mathscr{F}(\theta)$ corresponds to the domain of the moment generating function of the reference distribution considered.

The main results of this section are given in the next three theorems. They show that there is a correspondence between the reference distribution with respect to which the maximum entropy prior $p(\phi|\theta)$ is derived and the likelihood one would obtain by setting

in (6)

$$p(\theta|d) = \prod_{i=1}^{n} \pi_i^{\gamma}(\theta)$$

From Ragusa (2005), the solution for $\pi_i^{\gamma}(\theta)$ is given by

$$\pi_i^{\gamma}(\theta) = \tilde{\gamma}_1(\zeta' q(w_i, \theta))$$

where $\tilde{\gamma}_1(\cdot)$ denotes the inverse function of $\gamma(\cdot)$ and ζ is the Lagrange multiplier relative to the constraint $\sum_{i=1}^{n} q(w_i, \theta)\pi_i = 0$. The characterization of the posterior in (10) allows to derive a correspondence between the reference distribution under which the minimum entropy prior is derived and the likelihood obtained by replacing $p(w^n|\theta)$ with $\prod_{i=1}^{n} \pi_i^{\gamma}$.

Theorem 3. *Suppose $p(\phi) = \prod_{i=1}^{n} \exp(-\phi_i)$, that is $\{\phi_1, \ldots, \phi_n\}$ are independently and identically distributed as $\mathcal{G}(1, 1)$. Then,*

$$p(w^n|\theta) \propto \prod_{i=1}^{n} \varphi_1(\phi_i)\ p(\theta) = \prod_{i=1}^{n} \frac{1}{1 - \zeta' q(w_i, \theta)}\ p(\theta)$$

and

$$\zeta = \arg \min_{\zeta \in \mathcal{Z}(\theta)} \sum_{i=1}^{n} -\log(1 - \zeta' q(w_i, \theta))$$

where $\mathcal{Z}(\theta) = \{\zeta : \zeta' q(w_i, \theta) < 1, i = 1, \ldots, n\}$.

Theorem 3 provides a probabilistic interpretation to the idea of using the profiled empirical likelihood as a marginal likelihood. However, by choosing carefully the reference prior $p(\phi)$ other profiled likelihoods corresponding to minimum divergence solutions can be obtained.

Theorem 4. *Suppose $p(\phi) = \prod_{i=1}^{n} \phi_i \exp(-\phi_i)$, that is $\{\phi_1, \ldots, \phi_n\}$ are independently and identically distributed as $\mathcal{P}(1)$. Then,*

$$p(w^n|\theta) \propto \prod_{i=1}^{n} \varphi_1(\phi_i)\ p(\theta) = \prod_{i=1}^{n} \exp\{\zeta' q(w_i, \theta)\}\ p(\theta)$$

and

$$\zeta = \arg \min_{\zeta \in \mathcal{Z}(\theta)} \sum_{i=1}^{n} \exp(\zeta' q(w_i, \theta))$$

where $\mathcal{Z}(\theta) = \{\zeta : -\infty < \zeta < \infty\}$.

The following result extends to the continuous updating estimator.

Theorem 5. *Suppose* $p(\phi) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-\phi_i^2)$*, that is* $\{\phi_1, \ldots, \phi_n\}$ *are independently and identically distributed as* $\mathcal{N}(0,1)$*. Then,*

$$p(w^n|\theta) \propto \sum_{i=1}^{n} \varphi_1(\phi_i)\ p(\theta) = \sum_{i=1}^{n} (1 + \zeta' q(w_i, \theta))\ p(\theta)$$

and

$$\zeta = \arg \min_{\zeta \in \mathcal{Z}(\theta)} \sum_{i=1}^{n} \left\{ \left[\zeta' q(w_i, \theta))\right]^2 - \left[\zeta' q(w_i, \theta)\right] \right\}$$

where $\mathcal{Z}(\theta) = \{\zeta : -\infty < \zeta < \infty\}$.

Since $p(w^n|\theta)$ represents the joint probability of w^n, we must have that $\sum_i \varphi_1(\lambda' q_i(\theta)) = 1$. In general the weights would not sum to 1, except in the well known case of EL. In general the weights will be normalized to 1

$$\prod_{i=1}^{n} \frac{\varphi_1(\lambda' q_i(\theta))}{\sum \varphi_1(\lambda' q_i(\theta))}$$

Rescaling is hardly restrictive and it can be shown to be consistent with a mean shift of the initial prior under which the prior is derived. Theorem 4 and 5 give likelihoods that are equivalent to those obtained by using the weights of Empirical Tilting and Continuous Updating, that is the weights obtained from the MD problem with $\gamma(x) = \exp(x) - x - 1$ and $\gamma(x) = 1/2x^2 - x$. Notice that the likelihood derived in Theorem 5 may be improper, since its components can take negative values. The possibility of getting a negative posterior is due to the fact that the prior used in Theorem 5 for obtaining the likelihood is improper, giving positive probability to negative values of components of ϕ. The case of a negative posterior is the Bayesian version of the fact that in the minimum divergence problem for the CUE the optimal weights can be negative.

Ragusa (2005) considers the Hyperbolic Tilting as an alternative to the the empirical likelihood in a classical framework. The weights of the hyperbolic tilting are given by $\pi_i = \exp\{\sinh(\zeta' q(w_i, \theta))\} \cosh(\zeta' q(w_i, \theta))$. Although it is not possible to characterize explicitly the reference prior distribution to obtain the weights of the hyperbolic tilting, the reference prior must satisfy $\phi_1(\tau) = \exp\{\sinh(\tau)\} \cosh(\tau)$. In other cases it is possible to obtain the likelihood starting from the prior. Assuming the reference prior given by the density $p(x) = \exp(-\exp(-x)) \exp(-x)$ results in a likelihood where the components are given by $\Gamma'(1-x)/\Gamma(1-x)$, for $x > 1$, the Digamma function. It is also feasible to consider mixture of priors. Using the results in Theorem 3 and 4, the likelihood that corresponds to the prior $p(\phi) = \alpha \{\prod \exp(-\phi_i)\} + (1 - \alpha) \{\prod \phi_i \exp(-\phi_i)\}$, $\alpha \in (0, 1)$ is easily seen to

be

$$\prod_{i=1}^{n} \alpha \frac{1}{1-\zeta' q(w_i,\theta)} + (1-\alpha) \prod_{i=1}^{n} \exp(\zeta' q(w_i,\lambda))$$

where ζ solves

$$\zeta = \arg \min_{\zeta \in \mathcal{Z}(\theta)} \alpha \sum_{i=1}^{n} \exp(\zeta' q(w_i,\theta)) + (1-\alpha) \sum_{i=1}^{n} -\log(1-\zeta' q(w_i,\theta))$$

where $\mathcal{Z}(\theta) = \{\zeta : \zeta' q(w_i,\theta) < 1\}$.

Figure 3(a) illustrates the (density) prior that gives empirical likelihood and the (probability mass) that gives exponential tilting. In Figure 3(b) the shapes of the optimal weights are plotted for the two methods. The two priors treat differently the event $\phi = 0$. The prior for exponential tilting is well defined at $\phi = 0$, whereas $\phi = 0$ is outside the support for the empirical likelihood reference prior. When zero is excluded from the support of the reference prior, the set of values of $\lambda' q(w_i,\theta)$ for which the components of the likelihood are defined is a subset of the real line. This interpretation parallels the claims in Schennach (2003) and Ragusa (2005) regarding the inability of EL to deliver meaningful estimators when misspecification is present. We conjecture that estimators that are well defined when the model condition is misspecified must have weights that can be obtained through a reference prior whose support includes zero.

4 Validity of the Posterior

Monahan and Boos (1992) question the validity of a Bayesian analysis whenever the proposed likelihood is not precisely the conditional density $f(d|\theta)$ of the data given θ. Valid posterior probability statements are those that are supported by probability calculus. By "probability calculus" we mean here application of the Bayes theorem. The analysis of the previous section is based on the Bayes theorem, but it may be disputed on the ground that the maximum entropy prior is data-dependent. Hence, it is interesting to see if the likelihoods obtained by divergence weights are valid posterior distributions. Monahan and Boos (1992) suggests a definition of validity based on the coverage properties of posterior sets and they also provide a simple numerical technique that may be used to invalidate certain likelihoods.

Definition 1 (Coverage Set)**.** A coverage set function $R_\alpha(d)$ is a coverage set function of level α for a prior $p(\theta)$ and a likelihood $\ell(d|\theta)$ if for every d $\Pr\{\theta \in R_\alpha(w^n)\} = \alpha$ under the conditional measure $p_*(\theta|d) = p(\theta)\ell(d|\theta)$.

In the simple one-dimensional case, the natural posterior coverage set functions are

one-sided intervals $R(w) = (-\infty, t_\alpha]$, where

$$\int_{-\infty}^{t_\alpha} \ell(d|\theta)p(\theta)\mathrm{d}\theta = \alpha$$

Definition 2 (Valid posterior by coverage)**.** A posterior is said to be valid if for every $R_\alpha(d)$, $\Pr\{\theta \in R_\alpha(w)\} = \alpha$ under the joint measure $p(\theta, d) = p(\theta)f(d|\theta)$ on (θ, d).

If the likelihood is based on $f(d|\theta)$ the posterior is correct by coverage, since

$$\begin{aligned}
& \int\int 1(t \in R_\alpha(s))p(t,s)dtds \\
= & \int\int 1(t \in R_\alpha(s))p(t)p(s|t)dtds \\
= & \int\int 1(t \in R_\alpha(s))p(t)\ell(s|t)dt \left(\int p(t)p(s|t)dt\right) ds \\
= & \alpha \int\int p(t)p(s|t)dtds \\
= & \alpha
\end{aligned}$$

A likelihood $\ell(w|\theta)$ is proper Bayesian likelihood if and only if for every absolutely continuous prior $p(\theta)$, the posterior $\ell(w|\theta)p(\theta)$ is valid by coverage. Simply put, validity of the posterior means that all the intervals above have correct coverage. Let

$$H = \int_{-\infty}^{\theta} \ell(d|\theta)p(\theta)\mathrm{d}\theta$$

If the posterior is valid by coverage, then H is Uniform (0,1) under $p(\theta)f(d|\theta)$. The random variable H can be used to invalidate the use of a particular likelihood function in Bayesian analysis. The method works as follows:

1. generate θ_j, $j = 1, \ldots, u$ independently from the continuous distribution $p(\theta)$;
2. for each θ_j generate w^n from $f(w|\theta_j)$;
3. compute $H_j = \int_{-\infty}^{\theta_k} \ell(d|t)p(t)\mathrm{d}t$.

Manahan and Boos (1992) propose to test whether the sample comes from a Uniform (0,1) by using a Kolmogorov-Smirnov test. An alternative is to consider quantile-quantile plots comparing the quantiles of H_j, $J = 1, \ldots, u$ and the quantiles of the Uniform (0,1) distribution. This method is particular convenient because the Kolmogorov-Smirnov test may be sensitive to the approximation error due to the numerical integration error.

Using this technique we investigate a simple example to verify if likelihoods based on EL, ET and HT are valid posteriors by coverage. We consider two different experiments. In

both experiments, at each replications, $\tilde{\theta}$ is drawn from $p(\theta)$ and $w_1, \ldots, w_n$ are generated from $N(\tilde{\theta}, 1)$ and the following moment equation is considered

$$q(w, \theta) = \begin{bmatrix} w - \theta \\ (w - \theta)^2 - 1 \end{bmatrix}$$

In the first experiments $p(\theta)$ is Uniform$(0, 1)$; in the second $p(\theta)$ is Uniform$(-2, 2)$. The choice of a uniform prior is for convenience, because it the numerical integration required to rescale the likelihood can be carried on a finite interval. The EL, ET and HT likelihoods are calculated by solving their respective dual problem for λ for each θ. The numerical integrals are computed using an adaptive quadrature routine. Figure 4 and Figure 5 show quantile-quantile plots of the random variable H obtained by considering a Uniform$(0, 1)$ prior and likelihoods that are products of the EL, ET, HT weights respectively. Lazar (2003) finds that Empirical Likelihood gives valid likelihood with uniform prior at $n = 50$. The two figures extend this finding to other likelihoods, notably to BETEL of Schennach (2005). Figure 6 and Figure 7 show the quantile-quantile plots of H when a Uniform $(-2, 2)$ prior is considered. Inspection of the figures reveals that a considerable increase in the variance of the prior does not lead to a noticeable departure of H from uniformity.

5 Asymptotic Properties of MD/GEL Posteriors

5.1 Maximum Posterior Probability Estimator

As already pointed out, the relationship between MD/GEL is only perfect for the Empirical Likelihood, that corresponds to the case in which the prior distribution for the nuisance parameters is taken to be $Exp(1)$. For the other cases, the MD/GEL estimator do not coincide with the Maximum Posterior Probability (MPP) estimator (when $p(\theta) \propto 1$). However, under suitable conditions the MPP estimator is $\sqrt{n}$ consistent estimator of θ_0. General assumptions, equivalent to those of Theorem 8, could be given under which the (MPP) estimator is consistent. For reasons that would become clear later, we study the asymptotic behavior of the estimators under more standard regularity conditions.

Assumption 1.

(i) $\theta_0 \in \text{Int}(\Theta)$, *where* Θ *is convex compact subset of* $\mathbb{R}^k$;

(ii) $E\left[q(W, \theta_0)\right] = 0$ *and* $\|E\left[q(W, \theta)\right]\| \neq 0$, $\theta \neq \theta_0$;

(iii) $\sup_{\theta \in \Theta} E\|q(w, \theta)\|^2 < \infty$

(iv) $\sup_{\theta \in \Theta} E\|\partial q(w, \theta)/\partial \theta\|^2 < \infty$

(v) $q(w,\theta)$ *is continuous in* θ *at each* $\theta \in \Theta$*;*

(vi) $q(w,\theta)$ *is continuously differentiable at each* $\theta \in B_\delta(\theta_0)$.

Let $\hat{\theta} = \arg\max_{\theta\in\Theta} p_n(\theta|d)$. We are ready to state the following results.

Theorem 6. *If Assumption 1(i)-(vi) hold, then*

(i) $\sqrt{n}(\hat{\theta} - \theta_0) = o_p(n^{1/2})$

(ii) $J(\theta_0)^{-1}\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I_k)$

The priors that define the likelihoods are very difficult to interpret and hence to distinguish. For a given problem it is not clear which of the MD/GEL likelihood is appropriate for the problem. A possibility is to find a likelihood that has good higher order properties. For example, one could consider using only likelihoods that lead to maximum posteriors that are higher order efficient. This would lead to choose either the Empirical Likelihood or the BETEL based likelihoods, that have been proved to deliver higher order efficient estimators as in Newey and Smith (2005) and Schennach (2005).

Assumption 2. *There exists a function* $\mathcal{B}(w_i)$ *with* $E(\mathcal{B}(w_i))^6 < \infty$ *such that, in a neighborhood* θ_0 *, all the partial derivatives of* $q(w_i, \theta_0)$ *with respect to* θ_0 *up to order 4 exist, are bounded by* $\mathcal{B}(w_i)$ *and are Lipschitz in* θ *with prefactor* $\mathcal{B}(w_i)$*. The function* $\varphi(\cdot)$ *is five times continuously differentiable in a neighborhood of* 0.

Theorem 7. *Suppose Assumption 1-2 hold. Then the maximum posterior estimators,* $\hat{\theta}_{MPP}$*, defined as*

$$\hat{\theta}_{MPP} = \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} \frac{\varphi_1(\lambda' q_i(\theta))}{\sum_i \varphi_1(\lambda' q_i(\theta))}$$

admits a $O_p(n^{-2})$ *expansion and has the same* n^{-2} *variance of the maximum Empirical Likelihood estimator.*

5.2 Asymptotic for MD/GEL posterior

In a parametric setting, it is a well known fact that, under regularity conditions, parametric posterior distributions converge, as $n \to \infty$, to normal distributions. The idea can be sketched heuristically. Writing the posterior as

$$\begin{aligned} p(\theta|d) &\propto \prod_{i=1}^{n} p(w_i|\theta)p(\theta) \\ &\propto \exp\left\{\log p(\theta) + \sum_{i=1}^{n} \log p(w_i|\theta)\right\} \end{aligned}$$

The two logarithmic terms can be expanded about their respective maxima, π_0 and $\hat{\theta} \equiv \arg\max \log \sum_i p(w_i|\theta)$. Noting that $\partial \log p(\theta)/\partial\theta = 0$ and $\partial \log p(w_i|\theta)/\partial\theta = 0$, and setting $H_0(\theta) = \partial^2 \log p(\theta)/\partial\theta\partial\theta$ and $H(\theta) = \sum_i \partial^2 \log p(w_i|\theta)/\partial\theta\partial\theta$, the terms expand as follows

$$\begin{aligned} \log p(\theta) &= \log p(\pi_0) - (\theta - \pi_0)' H_0(\pi_0)(\theta - \pi_0)/2 + R_0 \\ \log \sum_{i=1}^{n} p(w_i|\theta) &= \sum_{i=1}^{n} \log p(w_i|\hat{\theta}) - (\theta - \hat{\theta})' H(\hat{\theta})(\theta - \hat{\theta})/2 + R_n \end{aligned}$$

where R_n and R_0 are remainders terms, which under regularity conditions are small for large n. Substituting the expansions in the posterior and ignoring normalizing constants, we obtain

$$\begin{aligned} p(\theta|d) &\propto \exp\left\{-(\theta - \pi_0)' H_0(\pi_0)(\theta - \pi_0)/2 - (\theta - \hat{\theta})' H(\hat{\theta})(\theta - \hat{\theta})/2\right\} \\ &\propto \exp\left\{-(\theta - m_n)' H_n (\theta - m_n)/2\right\} \end{aligned}$$

where $H_n = H_0(\pi_0) + H(\hat{\theta})$ and $m_n = H_n^{-1}\left(H_0\pi_0 + H(\hat{\theta})\hat{\theta}\right)$. The above expansion suggests that $p(\theta|d)$ will asymptotically approach a multivariate normal distribution with mean m_n, the weighted average of the prior mode and the maximum likelihood estimate and variance H_n, the sum of the observed information matrix. By application of the law of large numbers, it is straightforward to see that $H(\hat{\theta}) \to nI(\hat{\theta})$, the expected information matrix.

In a general context, Ibragimov and Has'minskii (1981, Chapter 1) demonstrate that Bayes estimators converge almost surely to the true value of the parameter when the number of observations grow to infinity. This is the case for estimators ξ that minimize the posterior loss associated with the loss function $\rho(\theta - \xi)$, under fairly weak conditions on the prior distribution and the sampling density. Ibragimov and Has'minskii (1981, Chapter 3) also prove the asymptotic efficiency of some Bayes estimates, that is, that the posterior distribution converges to the true value at rate $n^{-1/2}$.

The mentioned papers and the heuristic above are intended for parametric likelihood. Nonetheless, Chernozhukov and Hong (2003) show that is possible to extend the theory in Ibragimov and Has'minskii (1981) to a semi-parametric setting. They consider Laplacian quasi-posterior defined as

$$\mathcal{L}_n(\theta) = \frac{\exp\{L_n(\theta)\}p(\theta)}{\int \exp\{L_n(\theta)\}p(\theta)\mathrm{d}\theta}$$

and relative Bayes estimator defined as

$$\arg\inf \sqrt{n} \int \rho(\theta - \xi)\mathcal{L}_n(\theta)\mathrm{d}\theta$$

where $L_n(\theta)$ is a criterion function. Here, $\mathcal{L}_n(\theta)$ plays the same role than a posterior distribution, but there is no guarantee that $\mathcal{L}_n(\theta)$ is a proper posterior distribution. Under suitable conditions on $L_n(\theta)$ they prove that $\mathcal{L}_n(\theta)$ is concentrated at a $1/\sqrt{n}$ neighborhood of θ_0, and in this neighborhood $\mathcal{L}_n(\theta)$ is approximately a random normal density.

Chernozhukov and Hong (2003) also consider different specifications for the criterion function. In particular, when the parameter of interest is defined implicitly through a moment condition, $E[q(W,\theta_0)] = 0$, they consider both a GMM criterion and a GEL criterion function

$$L_n(\theta) = \sum_{i=1}^{n} s(\lambda(\theta)' q(w_i,\theta)), \quad \text{where} \quad \lambda(\theta) = \arg \inf_{\lambda \in \Lambda_n(\theta)} \sum s(\lambda' q(w_i,\theta))$$

for a strictly convex function $s(\cdot)$ defined in a neighborhood $\mathcal{A}$ of 0, and $\Lambda_n(\theta) = \{\lambda : \lambda' q(w_i,\theta) \in \mathcal{A}\}$.

The results of Chernozhukov and Hong (2003) can be readily extended to the posterior distributions derived in Section 3,

$$p_n(\theta|d) \quad = \quad \frac{\prod_{i=1}^{n} \varphi_1(\lambda' q(w_i,\theta)) p(\theta)}{\int \prod_{i=1}^{n} \varphi_1(\lambda' q(w_i,\theta)) p(\theta) \mathrm{d}\theta}$$

Let $H_n = \sqrt{n}(\theta - \theta_0) - J(\theta_0)^{-1}\Delta_n(\theta_0)/\sqrt{n}$, $J(\theta_0) = G(\theta_0)' V^{-1} G(\theta_0)$ and $\Delta_n(\theta_0)/\sqrt{n} = \sum_i q(w_i,\theta)'/\sqrt{n} V(\theta_0)^{-1} G(\theta_0)$, where $G(\theta) = \nabla_\theta E[q(w,\theta)]$. Also,

$$p_n^*(h|d) \quad = \quad p_n\left(h/\sqrt{n} + \theta_0 + J(\theta_0)^{-1}\Delta_n(\theta_0)/n\right)/\sqrt{n}$$

and $h = \sqrt{n}(\theta - \theta_0) - J(\theta_0)\Delta_n(\theta_0)/n$. Convergence in probability and distribution are under the outer probability, P^*.

Assumption A.

(i) $\theta_0 \in \mathrm{Int}(\Theta)$, *where* Θ *is convex compact subset of* $\mathbb{R}^k$;

(ii) the penalty function $\rho(\cdot)$ *is convex,* $\rho(h) \geq 0$, *with equality holding if and only if* $h = 0$, $\rho(h) \leq 1 + |h|^p$ *for some* $p \geq 1$, *and* $\int_{\mathbb{R}^k} \rho(u - \xi)\exp(-u'au)\mathrm{d}u$ *is uniquely minimized at some point* $\xi \in \mathbb{R}^k$.

(iii) $E\left[q(W,\theta_0)\right] = 0$ *and* $\|E\left[q(W,\theta)\right]\| \neq 0$, $\theta \neq \theta_0$;

(iv) $\partial P\left(q(W,\theta) < w\right)\partial\theta$ *is continuous in* θ *uniformly in* $w : \|w\| \leq K$, *for* K *in (iii);*

(v) $\sup_{|\theta-\theta_0|<\delta} \|q(w,\theta)\| < K$, *a.s., for some constant* K;

(vi) $\{q(w,\theta), \theta \in \Theta\}$ *is Donsker class, where*

$$\sum_{i=1}^{n} q(w_i,\theta)/\sqrt{n} \xrightarrow{d} N(0,V), \qquad V \equiv E\left[q(w,\theta)q(w,\theta)'\right] > 0$$

Theorem 8. *If Assumption A(i)-(vi) hold:*

(I) As $n \to \infty$, $p_n(\theta)$ *is approximately normal as measured by total variation norm, that is*

$$\int_{H_n} |p_n(h|d) - p_\infty(h)| \mathrm{d}h \xrightarrow{p} 0$$

where

$$p^*_\infty(h) = \sqrt{\frac{|J(\theta_0)|}{(2\pi)^k}} \exp\left(-h'J(\theta_0)h/2\right)$$

(II) $J(\theta_0)^{-1}\sqrt{n}(\hat{\theta}_\rho - \theta) \xrightarrow{d} N(0, I_k)$

(III) Let $c_{g,n}(\alpha) = \inf\left\{x : \int_{\theta\in\Theta: g(\theta)\le x} p_n(\theta)\mathrm{d}\theta \ge \alpha\right\}$, $\alpha \in (0,1)$. *For every differentiable function* $g(\theta)$,

$$\lim_{n\to\infty} P^*\left\{c_{g,n}(\alpha/2) \le g(\theta_0) \le c_{g,n}(1-\alpha/2)\right\} = 1-\alpha$$

The proof of Theorem 8 follows from the proof of Proposition 2 in Chernozhukov and Hong (2003), and it is not surprising given the analysis of the extremum class of estimators defined in the previous section.

Theorem 8 implies that selected moments of the posterior MD/GEL can be interpreted from a classical perspective, being consistent estimators of θ_0. The marginal posterior distribution can also be used in the construction of asymptotically valid confidence intervals, establishing an asymptotic correspondence between High Posterior Density intervals and confidence bands.

6 Monte Carlo Markov Chain

Calculating the posterior density at a given point θ is a relatively simple computational task even if a nonlinear optimization step must be performed to obtain the parameter $\lambda(\theta)$. On the contrary, integrating over regions Θ can prove a formidable task. Markov Chain Monte Carlo (MCMC) is a popular computational method for generating samples from virtually any distribution p defined on a space $\mathcal{X}$. These samples are often used to efficiently compute expectations by invoking some form of the law of large numbers. Although designed for

parametric posterior distributions, MCMC, and the Metropolis-Hasting (MH) algorithm in particular, can be used for obtaining moments of the GEL/MD posterior.

A detailed explanation of the Metropolis-Hasting (MH) algorithm and its properties can be found in Robert and Casella (2005). Here, we explore the feasibility of the MH algorithm in this sitting mainly to shed light on practical differences with the approach based on the Bayesian bootstrap. We use a Random Walk implementation of the MH algorithm (MHRW). Basically, given the posterior density

$$p(\theta|d) \propto \prod_{i=1}^{n} \frac{\varphi_1(\lambda(\theta)'q(w_i,\theta))}{\sum_i \varphi_1(\lambda(\theta)'q(w_i,\theta))}$$

the MHRW generate $\left(\theta^{(1)},\ldots,\theta^{(\ell)}\right)$ following the scheme:

Let π be a symmetric distribution, that is $\pi(\eta) = \pi(-\eta)$. Given $\theta^{(t)}$, $t = 1,\ldots,\ell$

1. Generate $\eta_t \sim \pi(|\eta - \theta^{(t)}|)$
2. Take
$$\theta^{(t+1)} = \begin{cases} \eta_t & \text{with probability } \min\left\{1, \frac{p(\eta_t|d)}{p(\theta^{(t)}|d)}\right\} \\ \theta^t & \text{otherwise} \end{cases}$$

By invoking the law of large number, the sample so obtained can be used to compute expectations with respect to $p(\theta|d)$:

$$\frac{1}{\ell}\sum_{t=1}^{\ell} g(\theta^{(t)}) \xrightarrow{p} \int_{\Theta} g(\theta)p(\theta|d)\mathrm{d}\theta$$

The advantage of MHRW over the standard HM algorithm is that it allows a local exploration of the neighborhood of the current value of the Markov chain, since the candidate distribution π depends on the current state of the chain, η_t. Markov chains generated from application of the MHRW algorithms are not in general uniform ergodic, but under regularity conditions they enjoy geometric ergodicity properties.

To investigate the properties of this MHRW, we compare the output of MCMC and the posterior evaluated on a grid. Two modelss are considered. The first model considered is an extension of the Hall and Horowitz model:

$$g(w,\theta) = r(w,\theta)\begin{pmatrix} 1 \\ w_2 \\ w_3 - 1 \\ \vdots \\ w_9 - 1 \end{pmatrix}$$

where $r(w,\theta) = \exp(-0.72-(w_1+w_2)\theta+3w_2)-1$, $(w_1, w_2) \sim N(.16, .8I_2)$ and $(w_3, \ldots, w_9)$ are iid χ^2_1. The second model is a Instrumental Quantile regression specified as follows

$$g(w,\theta) = (\tau - 1(y < x\theta))\, z$$

where $z = (z_1, \ldots, z_5) \sim N(0, I_5)$, $y = \varepsilon$, $x = \rho z = \eta$, $(\varepsilon, \eta) \sim N(0, \Sigma)$, $\Sigma = 0.9I_2$ and $\rho = \sqrt{0.3/(5-0.3)}$. Given a single realization of 200 observations from each of the two models, the likelihood is evaluated on a grid of 4,000 points. Then 10,000 draws per model are obtained by employing a MHRW algorithm, with a normal proposal density with standard deviation $\sigma = .3$ for the Hall and Horowitz model and $\sigma = 0.5$for the Instrumental Quantile model. Figure 7 and Figure 7 plot the grid based posterior and the histogram of the resulting chain for the two models. It can be seen that the histogram of $\{\theta\}_j$ produced by the MHRW closely approximate the true posterior distribution, even for the instrumental quantile model where the posterior is a step function.

We apply the MHRW algorithm described above to the estimation of the effect of injury compensations on time out of work. Meyer, Viscusi and Durbin (1995) collect data on work related injury claims filed by a random sample of workers in Kentucky. These claims are filed when a person becomes temporary unable to work but is expected to recover and return to work and are necessary to obtain weekly benefits. The amount a worker is entitled for is a function of previous earnings, with earnings above a certain level corresponding to a certain benefit. Kentuky raised the threshold for the maximum benefit from \$131 to \$217 per week on July 15, 1980. Between 1979 and 1980, workers in the high earning group ($x_1 = 1$), those earning more than \$131 per week, experienced an increase in the benefit, whereas for workers in the low earning group ($x_1 = 0$) the benefit level remained unchanged. The basic specification of work considered in Meyer et al. (1995) is

$$E(\log Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2$$

where Y is the duration of the work related injury, $x_2 = 1$ if the injury occurred after the benefit increase and $x_2 = 0$ otherwise. As in Chamberlain and Imbens (2003), we consider estimating the linear predictor coefficient at different quantiles of the distribution, both when duration is expressed in logarithm and in number of week:

$$\begin{aligned} Q_\tau(\log Y|X = x) &= \beta_0(\tau) + \beta_1(\tau)x_1 + \beta_2(\tau)x_2 + \beta_3(\tau)x_1 \cdot x_2, \quad \tau \in (0,1) \quad (13a) \\ Q_\tau(Y|X = x) &= \gamma_0(\tau) + \gamma_1(\tau)x_1 + \gamma_2(\tau)x_2 + \gamma_3(\tau)x_1 \cdot x_2, \quad \tau \in (0,1) \quad (13b) \end{aligned}$$

As discussed in Chamberlain and Imbens (2003) the discreetness of the data in both specification makes the asymptotic approximation poor, since the latter assumes that the

residuals have a continuous distribution in a neighborhood of the origin. On the other hand, methods based on the posterior should have no problems. We consider the case of the median, $\tau = 0.5$. The result for the logarithmic case are presented in Table 1.

The posterior distribution is based on the quantile moment conditions

$$\sum_{i=1}^{n} \left(\frac{1}{2} - 1 \left(Y_i \leq \beta_0(.5) + \beta_1(.5)x_{1,i} + \beta_2(.5)x_{2,i} + \beta_3(.5)x_{3,i} \right) \right) \begin{pmatrix} 1 \\ x_{1,i} \\ x_{2,i} \\ x_{1,i} \cdot x_{2,i} \end{pmatrix} = 0$$

in the first specification, and

$$\sum_{i=1}^{n} \left(\frac{1}{2} - 1 \left(\log Y_i \leq \gamma_0(.5) + \gamma_1(.5)x_{1,i} + \gamma_2(.5)x_{2,i} + \gamma_3(.5)x_{3,i} \right) \right) \begin{pmatrix} 1 \\ x_{1,i} \\ x_{2,i} \\ x_{1,i} \cdot x_{2,i} \end{pmatrix} = 0$$

in the second. In the simulations, the initial draw of the MCMC series is taken to be the ordinary least-square estimate.

Figure 8 and Figure 9 plot the MS/GEL posterior for the case in which $\varphi_1(u) = \exp(\sinh(u))\cosh(x)$. In other simulations, not reported here, the posterior based on empirical likelihood showed the same qualitative behavior. The acceptance rate of the chain was 45%. We performed a series of diagnostic tests all indicating that the chain has converged. The posterior mean and median of β_3 for the logarithmic specification were, respectively, -0.026 and -0.031. The highest posterior density interval was $(-0.275, 0.247)$. The posterior mean and median of γ_3 were, respectively, 0.063 and 0.048. The highest posterior density interval was $(-1.094, 1.257)$. A quick comparison with the posterior in Chamberlain and Imbens (2003, Figure 2) for the same parameter on similar sample shows that the MD/GEL based posterior is smoother than the equivalent posterior obtained by the Bayesian bootstrap.

7 Conclusions

In this paper we study how the use of generalized empirical likelihood in the Bayes theorem can be justified from a probabilistic point of view. We find that when the underlying random vector has a discrete distribution and all the possible values in the sample space have been observed the use of likelihoods based on MD/GEL weights can be justified. The procedure consists in putting a prior on the nuisance parameters –the multinomial weights– and integrating them out with respect to a distribution that satisfies the moment

conditions and has the largest entropy w.r.t. the initial prior. Different initial priors deliver different likelihoods that are proportional to the product of the weights that solve a specific MD/GEL problem. We also discuss the Bayesian validity of these likelihoods in a more general setting.

We then consider the asymptotic properties of both the maximum posterior estimators and the Bayesian estimators defined as moments of the posteriors. An interesting result is that the maximum posterior estimators all have the same n^{-2} variance.

The analysis can be extended in many interesting directions. First it would be interesting to consider the validity of likelihood in a time series context. Intuitively one could consider likelihoods that are proportional to the product of MD/GEL weights when the moment function is blocked as in Kitamura (1997). This would entail putting a prior on the blocks. Another interesting extension would be to consider limited information MD/GEL likelihoods based on conditional moment conditions as in Kitamura et al. (2004).

	Quantiles						
	.10	.20	.40	.50	.60	.70	.90
$\beta_0(\tau)$	.405 (.129)	. 916 (.066)	1.253 (.004)	1.253 (.028)	1.705 (.023)	1.872 (.040)	2.526 (.043)
$\beta_1(\tau)$	0 (.188)	0 (.096)	0 (.057)	0 (.041)	0 (.034)	0 (.058)	.148 (.070)
$\beta_2(\tau)$	0 (.129)	0 (.083)	0 (.052)	.251 (.065)	0 (.032)	.143 (.057)	.336 (.0870)
$\beta_3(\tau)$	0.00 (.188)	.336 (.121)	.251 (.0748)	0.201 (0.0764)	.167 (0.059)	.236 (.088)	.146 (.134)

Table 1: Quantile regression Coefficients for Log Duration, Kentucky

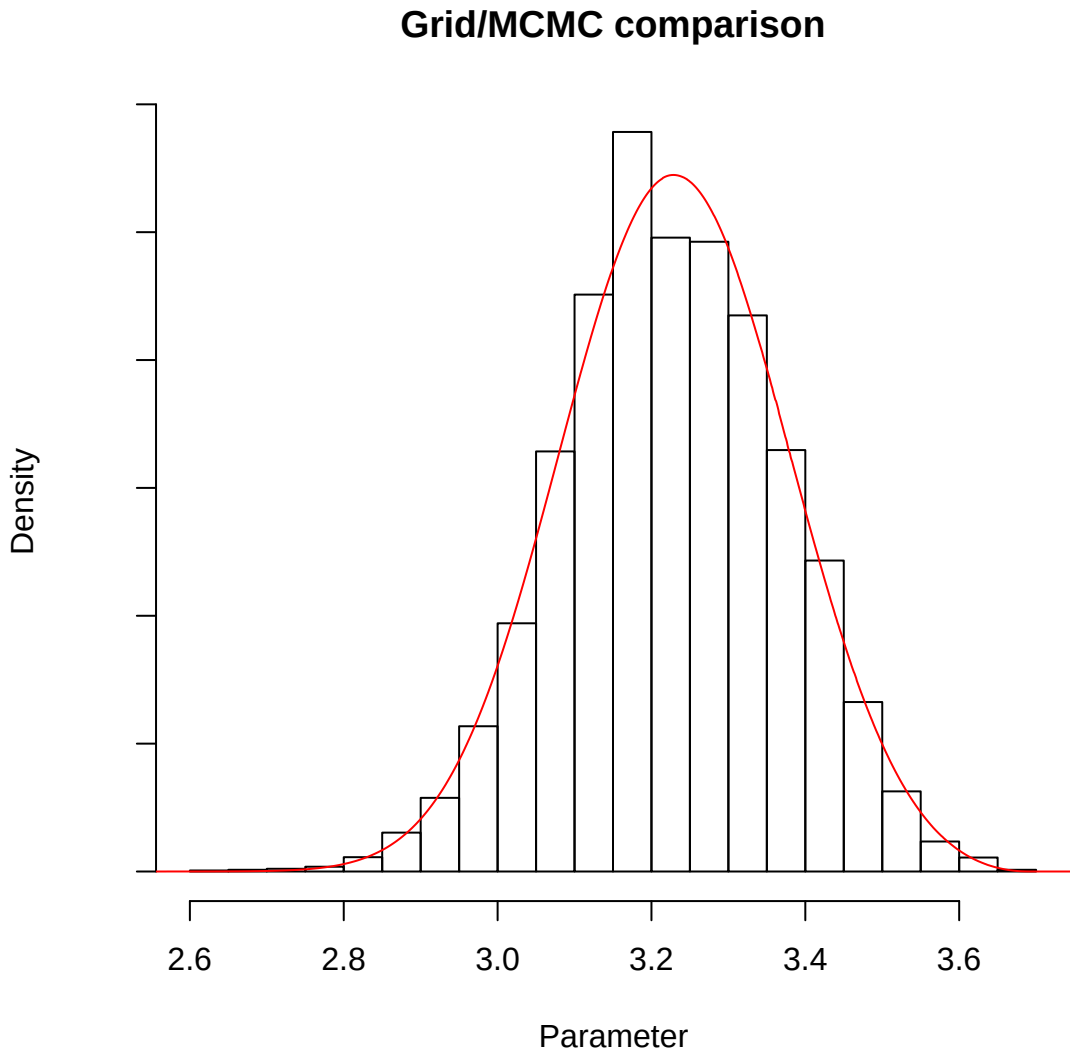

Figure 1: Comparison of posterior likelihood obtained by solving the dual problem on a grid (line) and histogram of a 10,000 draws from RWMH (histogram). Hall and Horowitz model.

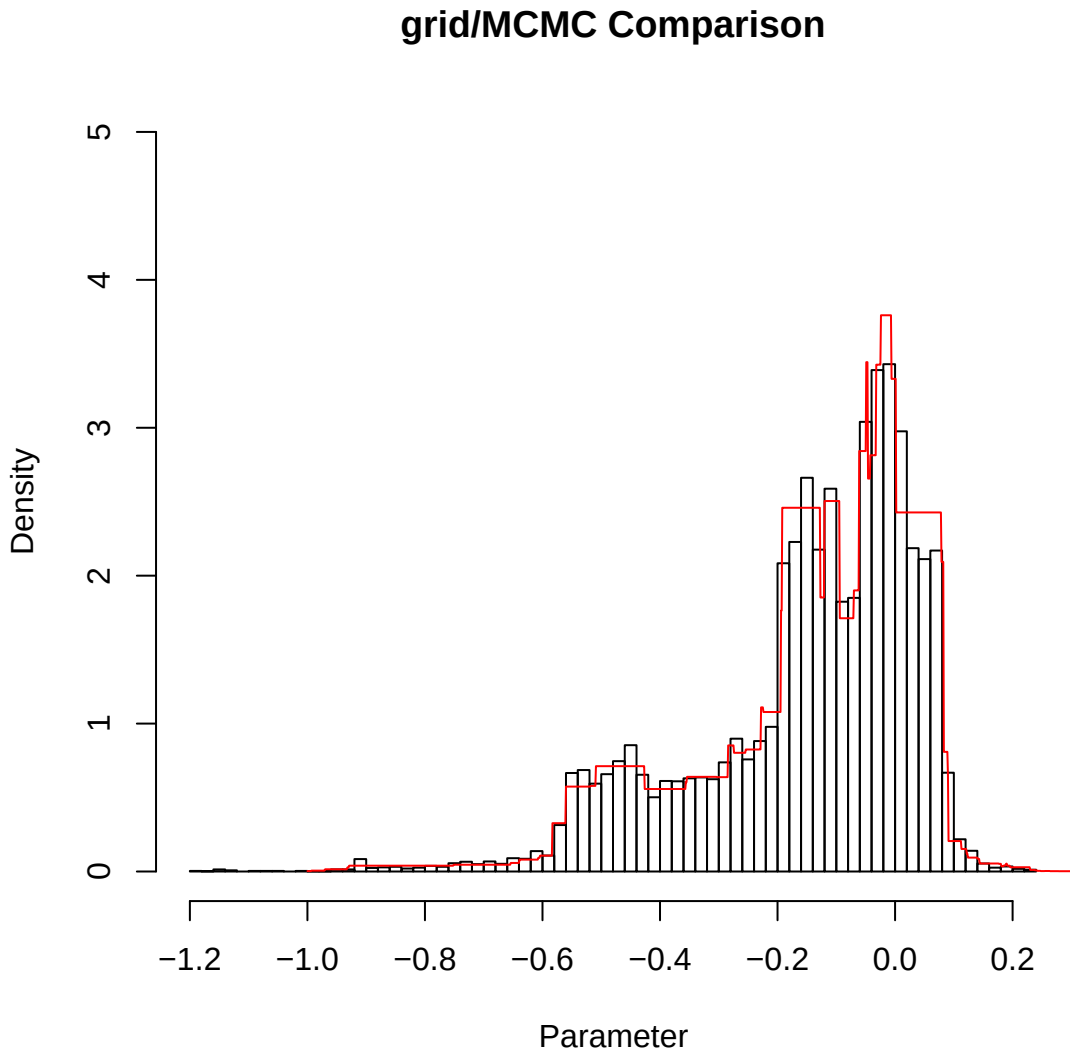

Figure 2: Comparison of posterior likelihood obtained by solving the dual problem on a grid (line) and histogram of a 10,000 draws from RWMH (histogram). Quantile instrumental variable model.

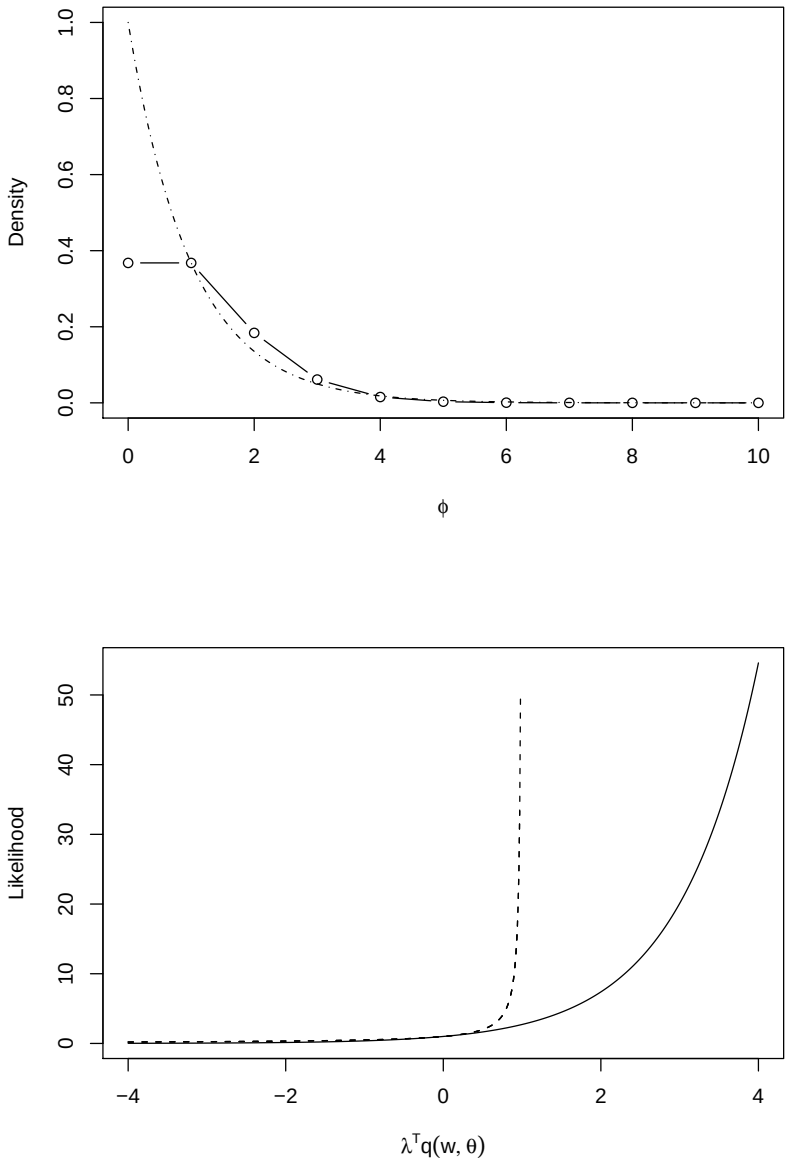

Figure 3: Reference Priors - Empirical Likelihood vs Exponential Tilting

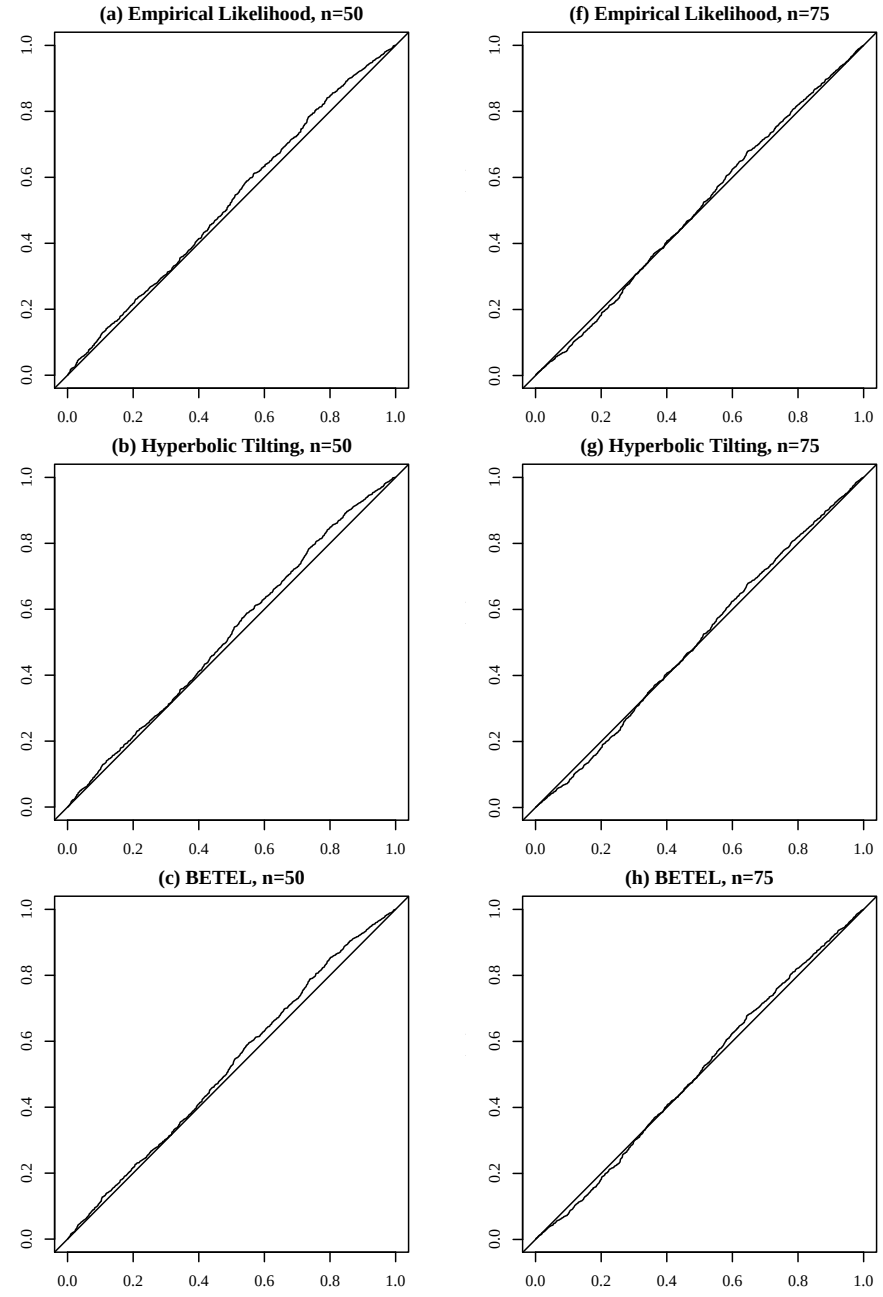

Figure 4: (a)-(f) show quantile-quantile plots for the distribution of H against Uniform$(0, 1)$, for the Normal Moment conditions, with uniform priors $(0, 1)$, sample sizes of $n = 50$ and $n = 75$ and simulation size of 1000.

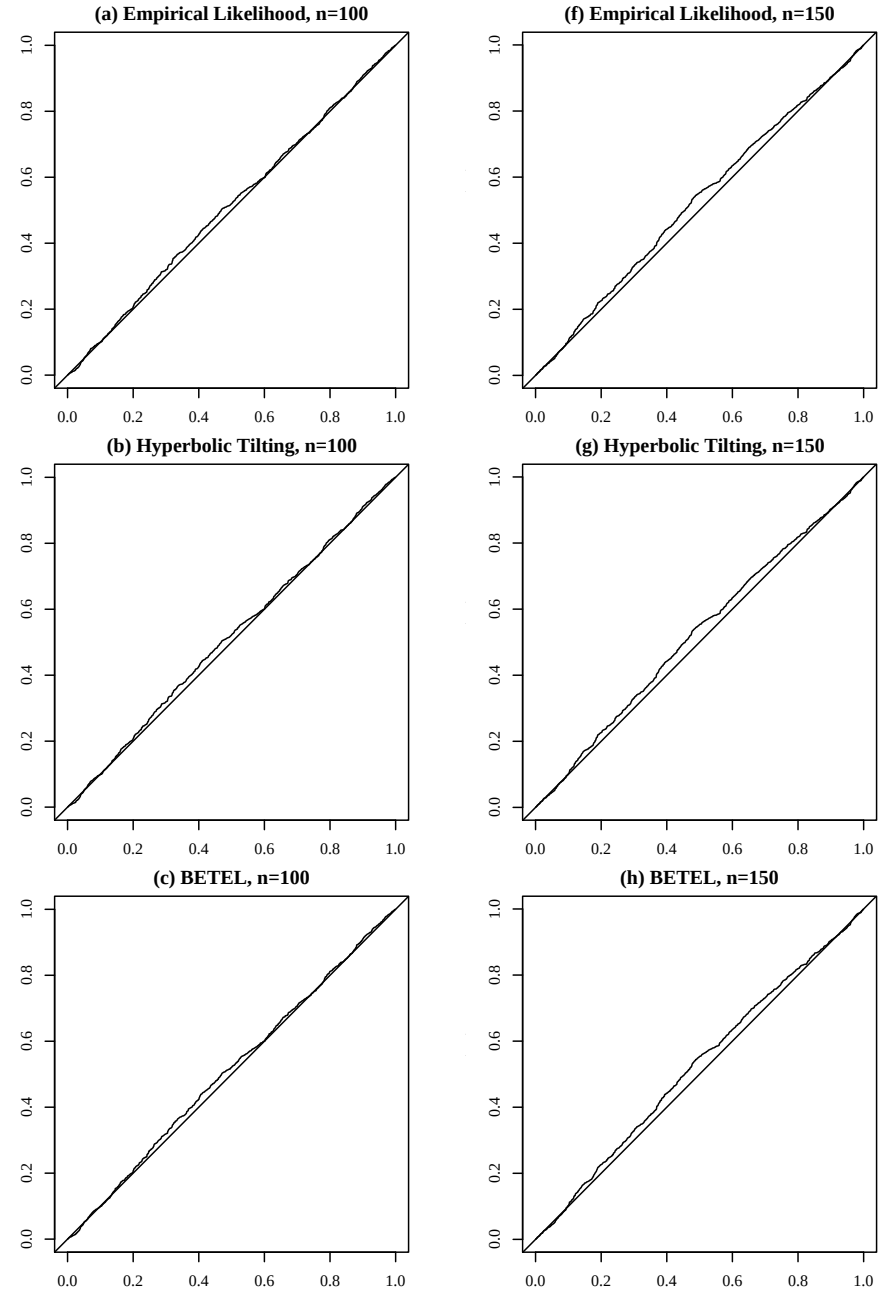

Figure 5: (a)-(f) show quantile-quantile plots for the distribution of H against Uniform$(0, 1)$, for the Normal Moment conditions, with uniform priors $(0, 1)$, sample sizes of $n = 100$ and $n = 150$ and simulation size of 1000.

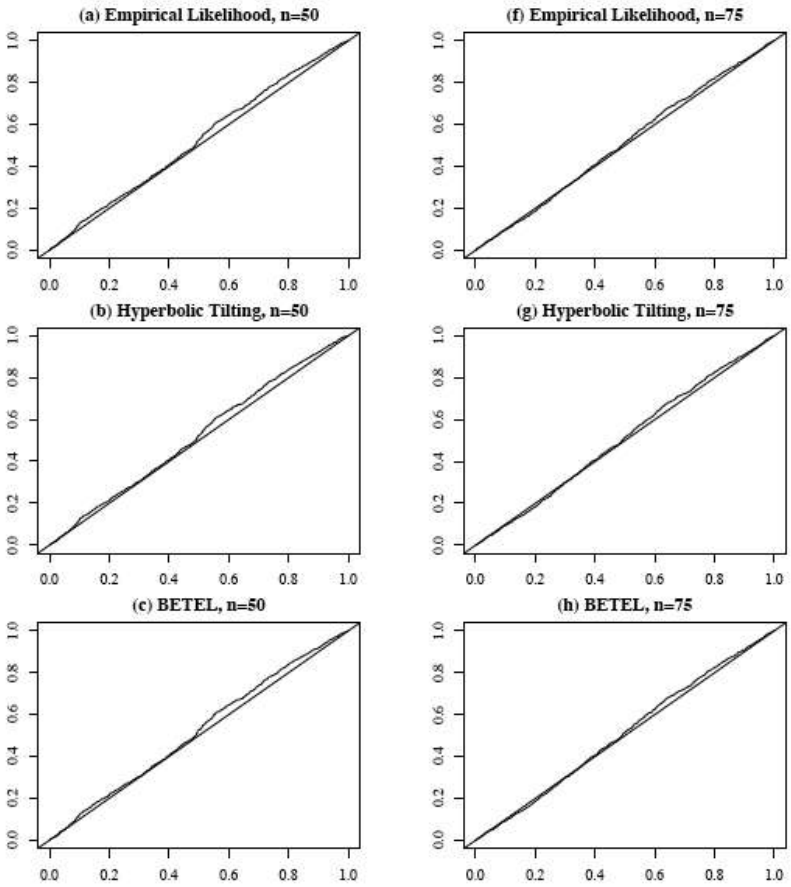

Figure 6: (a)-(f) show quantile-quantile plots for the distribution of H against Uniform$(0, 1)$, for the Normal Moment conditions, with uniform priors $(-2, 2)$, sample sizes of $n = 50$ and $n = 75$ and simulation size of 1000.

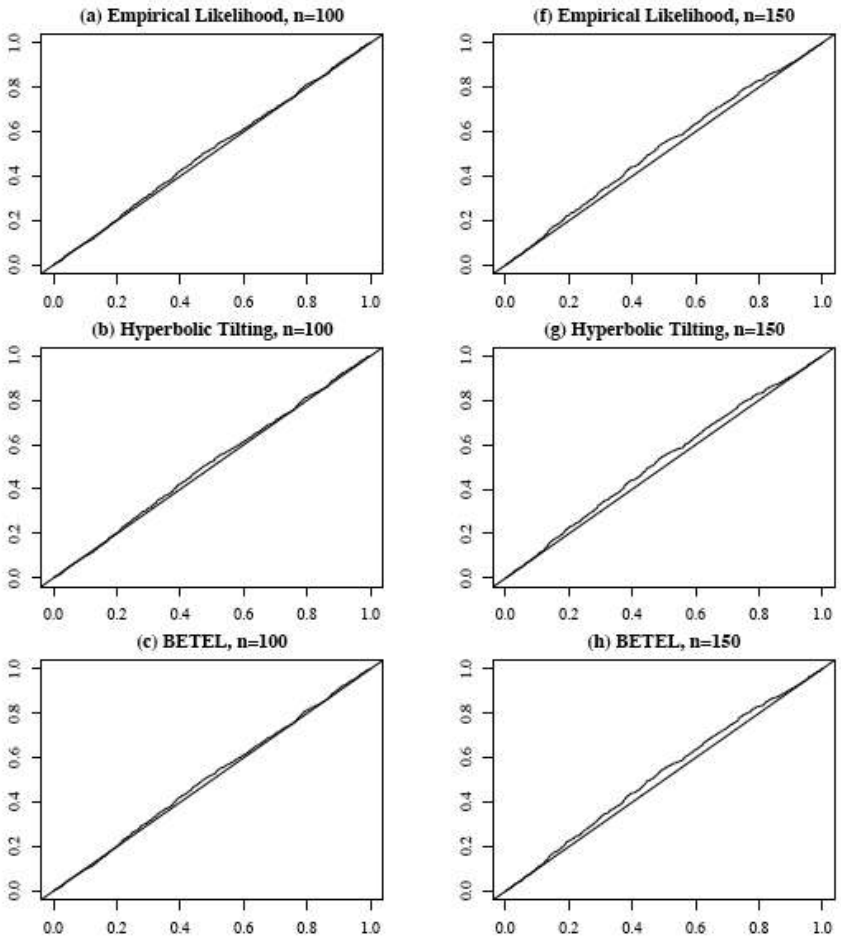

Figure 7: (a)-(f) show quantile-quantile plots for the distribution of H against Uniform$(0, 1)$, for the Normal Moment conditions, with uniform priors $(-2, 2)$, sample sizes of $n = 75$ and $n = 150$ and simulation size of 1000.

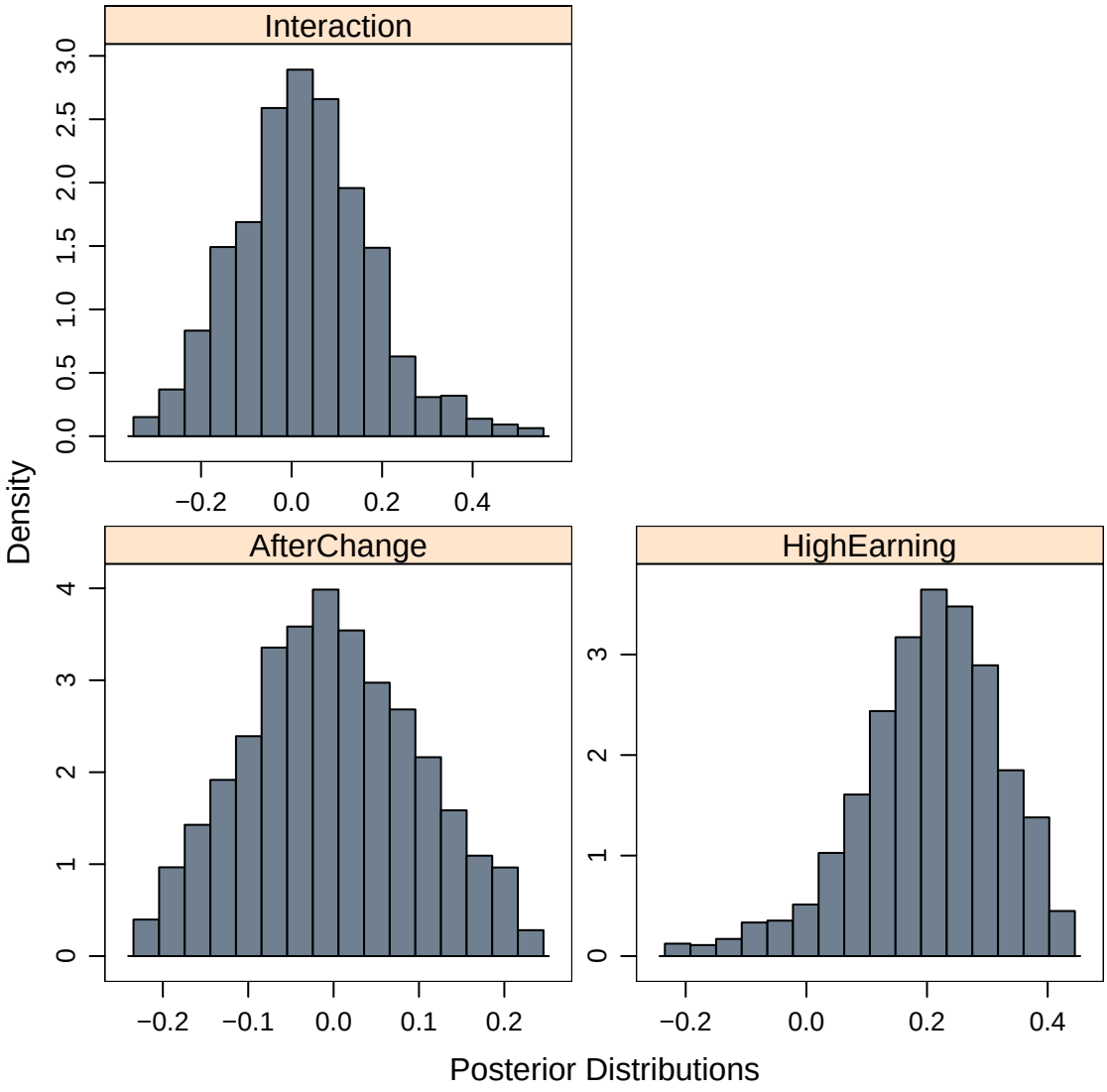

Figure 8: Posterior Histogram, $\tau = 0.5$, $\log(Duration)$ (HT)

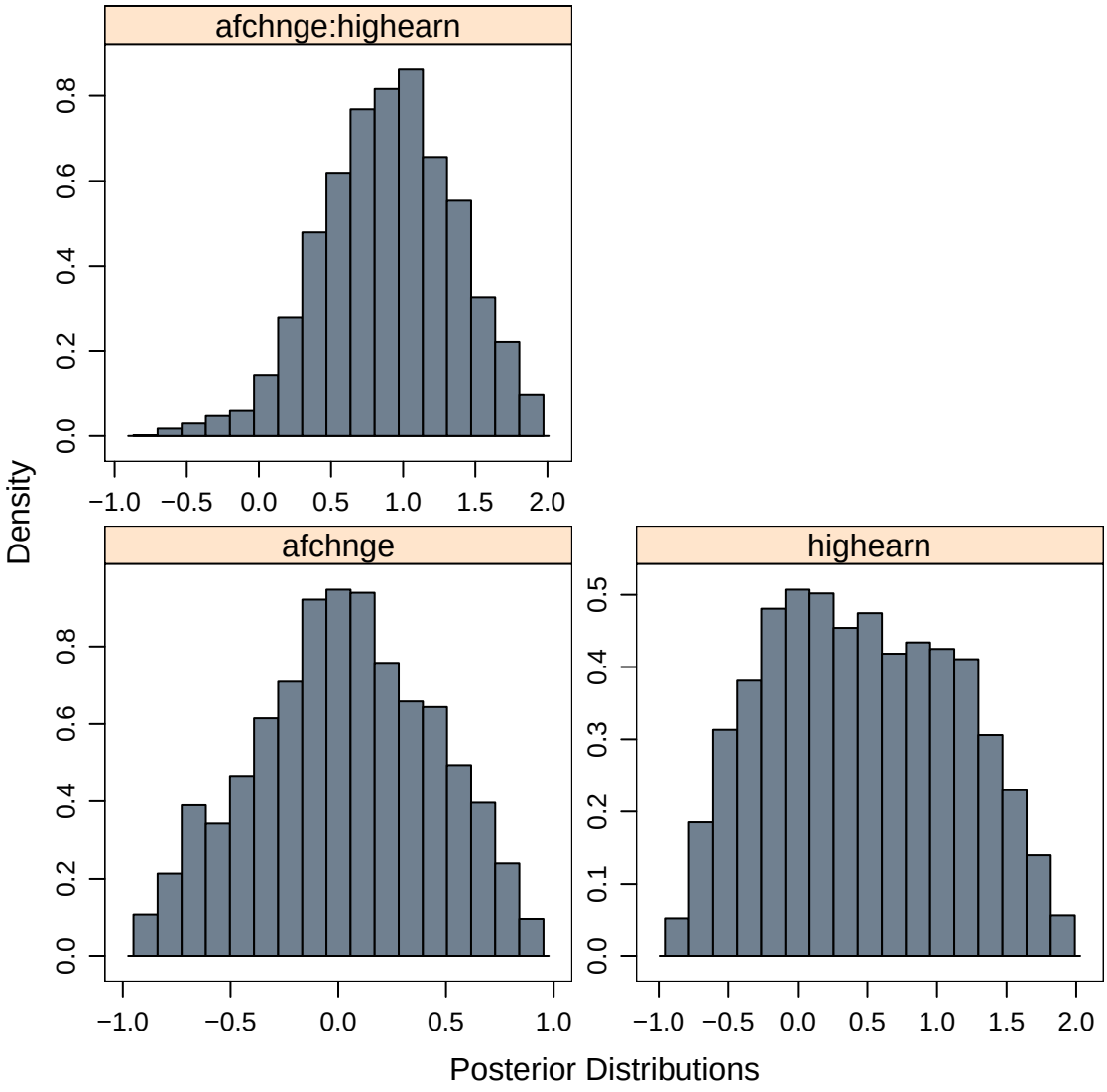

Figure 9: Posterior Histogram, $\tau = 0.5$, Duration in week, (HT)

A Mathematical Appendix

Proof to Theorem 1. Let $\sum_{i=1}^{n} q(w_i, \theta)\phi_i = \xi'\phi$. The Maximum Entropy problem is equivalent to the following minimization problem

$$\min_{P(\eta|\theta)} \int \log\left(\frac{dP(\eta|\theta)}{dP(\phi)}\right) dP(\eta|\theta)$$

s.t.

$$\int \xi'\phi dP(\phi|\theta) = 0$$

The solution to the following problem is well known to have the following Gibbs canonical density. Applying Theorem 3.1, Corollary 3.1, Theorem 3.3 of Csiszar(1975), we have

$$dP(\phi|\theta) = \frac{\exp\left\{\lambda' \sum_{i=1}^{n} q_i(\theta)\phi_i\right\} dP(\phi)}{\int \exp\left\{\lambda' \sum_{i=1}^{n} q_i(\theta)\phi_i\right\} p(\phi)d\phi} \tag{14}$$

where

$$\lambda = \arg\min_{\lambda} \int \exp\left\{\lambda' \sum_{i=1}^{n} q_i(\theta)\phi_i\right\} dP(\phi)$$

giving the desired result. □

Proof to Theorem 2. Rewrite the solution as

$$P(\phi|\theta) = \exp\left[\sum_{i=1}^{n} \lambda' q_i(\theta)\phi_i - \log \int \exp\left(\sum_{i=1}^{n} \lambda' q_i(\theta)\phi_i\right) p(\phi)\mathrm{d}\phi\right] P(\phi) \tag{15}$$

By the independence of $\{\phi_1, \phi_2, ..., \phi_n\}$ under $P(\phi)$, using the properties of the logarithm and of the exponential, and setting $\tau_i = \lambda' q_i(\theta)$, we have

$$\begin{aligned} \log \int \exp(\sum_{i=1}^{n} \tau_i\phi_i) p(\phi)\mathrm{d}\phi &= \log \prod_{i=1}^{n} \int \exp(\tau_i\phi_i) p(\phi_i)\mathrm{d}\phi_i \\ &= \sum_{i=1}^{n} \log \int \exp(\tau_i\phi_i) dP(\phi_i) \end{aligned}$$

Using the expression above, (15) can be rewritten as

$$P(\phi|\theta) = \exp\left[\sum_{i=1}^{n} [\tau_i\phi_i - \varphi(\tau_i)]\right] P(\phi) \tag{16}$$

and λ is now given by

$$\lambda = \arg \min_{\lambda \in \tilde{\Lambda}_n(\theta)} \sum_{i=1}^{n} \log \int \exp\left(\lambda' q_i(\theta)\phi_i\right) dP(\phi)$$

Independence follows from (16), since

$$\begin{aligned} P(\phi|\theta) &= \exp\left[\sum_{i=1}^{n}[\tau_i\phi_i - \varphi(\tau_i)]\right] P(\phi) \\ &= \prod_{i=1}^{n} \exp[\tau_i\phi_i - \varphi(\tau_i)]P(\phi_i) \end{aligned}$$

□

Proof to Theorems 3,4,5. We need to show that the moment generating function of Exponential, Poisson and Normal corresponds to the objective functions in the Theorems.

(1) The moment generating function of the Exponential distribution is given by $\int_0^{+\infty} e^{-x}e^{yx}\mathrm{d}x$ and hence in this case $\varphi(y) = \log \int_0^{+\infty} e^{-x}e^{yx}\mathrm{d}x = -\log(1-y)$ and $\varphi_1(y) = (1-y)^{-1}$.

(2) For the Poisson, the moment generating function is given by $\exp(\exp(y)-1)$ and hence $\varphi(y) = \exp(y) - 1$ and $\varphi_1(y) = \exp(y)$.

(3) The Moment generating function for a normal is given by $\exp(y+y^2/2)$ and $\varphi(y) = y + y^2/2$ and $\varphi_1(y) = 1 + y$. Combining the results in (1),(2) and (3) give the conclusion of the theorem.

□

Proof to Theorem 1. The MPP is defined as the maximum of the posterior distribution when $p(\theta) \propto 1$, that is

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} \frac{\varphi_1(\lambda(\theta)'q_i(\theta))}{\sum_{i=1}^{n}\varphi_1(\lambda(\theta)'q_i(\theta))} \\ &= \arg\min_{\theta\in\Theta} \left\{ \log \prod_{i=1}^{n} \frac{\varphi_1(\lambda(\theta)'q_i(\theta))}{\sum_{i=1}^{n}\varphi_1(\lambda(\theta)'q_i(\theta))} \right\} \\ &= \arg\min_{\theta\in\Theta} \left\{ \sum_{i=1}^{n} \log \varphi_1(\lambda(\theta)'q_i(\theta)) - \log \sum_{i=1}^{n} \varphi_1(\lambda(\theta)'q_i(\theta)) \right\} \end{aligned}$$

The proof follows along the line of Schennach (2003, Theorem 3), with the only difference that now the framework is more general. The first order conditions for $\hat{\theta}$ is given by

$$\left\{ \frac{\sum_{i=1}^{n}\varphi_2(\lambda(\theta)'q_i(\theta))}{\sum_{i=1}^{n}\varphi_1(\lambda'q_i(\theta))} \frac{\partial q_i(\theta)}{\partial\theta'}\lambda(\theta) - \sum_{i=1}^{n} \frac{\varphi_2(\lambda(\theta)'q_i(\theta))}{\varphi_1(\lambda(\theta)'q_i(\theta))} q_i(\theta)\frac{\partial\lambda(\theta)}{\partial\theta} \right\} = 0$$

whereas the first order conditions for $\hat{\lambda}$ is given by

$$\sum_{i=1}^{n} \varphi_1(\lambda' q_i(\theta)) q_i(\theta) = 0 \tag{17}$$

Since $\varphi_1(0) = \varphi_2(0) = 1$, expanding in a $O_p(n^{-1/2})$ neighborhood of $\theta = \theta_0$ and in $O(n^{-1/2})$ neighborhood of $\lambda(\theta_0) = 0$ the first order conditions gives the following expansions

$$\begin{pmatrix} 0 \\ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} q_i(\theta) \end{pmatrix} + \begin{bmatrix} 0 & \frac{1}{\sqrt{n}} \sum_i \frac{\partial q_i(\theta_0)}{\partial \theta} \\ \frac{1}{\sqrt{n}} \sum_i \frac{\partial q_i(\theta_0)}{\partial \theta'} & \frac{1}{\sqrt{n}} \sum_i q_i(\theta_0) q_i(\theta_0) \end{bmatrix} \begin{pmatrix} \theta - \theta_0 \\ \lambda \end{pmatrix} = o_p(1) \tag{18}$$

The GEL estimators, $\hat{\theta}_{GEL}$ and $\hat{\lambda}_{GEL}$ have the same asymptotic expansion and solve the same first order condition than the MPP. Since $\hat{\lambda}_{GEL} = o_p(1)$, and since the GEL objective function and the posterior distribution reaches its maximum values when $\hat{\lambda}_{GEL} = o_p(1)$ and $\hat{\lambda} = o_p(1)$, respectively, the existence of another solution outside the neighborhood of validity of expansion (18) can be ruled out. Hence, $\hat{\theta}_{GEL} = \hat{\theta} + o_p(1)$.

□

Proof of Theorem 7. First, as in Schennach (2006), it can be shown that $\hat{\theta}_{MPP} - \hat{\theta}_{EL} = O_p(n^{-3/2})$. It also follows that $\hat{\theta}_{MPP}$ admits an asymptotic expansion of the following form

$$\sqrt{n}(\hat{\theta}_{MPP} - \theta_0) = a(\theta_0) + b(\theta_0)/\sqrt{n} + c(\theta_0)/n\sqrt{n} + R_n(\theta_0)$$

where $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ are $O_p(1)$ smooth functions of moments of $q(w, \theta_0)$ and $R_n(\theta) = O_p(n^{-2})$. On the other hand, the EL estimator admits the following expansion

$$\sqrt{n}(\hat{\theta}_{EL} - \theta_0) = \tilde{a}(\theta_0) + \tilde{b}(\theta_0)/\sqrt{n} + \tilde{c}(\theta_0)/n\sqrt{n} + \tilde{R}_n(\theta_0)$$

The difference in the higher order variance is given by

$$\begin{aligned} & Var\left(\sqrt{n}(\hat{\theta}_{EL} - \theta_0)\right) - Var\left(\sqrt{n}(\hat{\theta}_{EL} - \theta_0)\right) \\ = \; & a(\theta_0)a(\theta_0)' - \tilde{a}(\theta_0)\tilde{a}(\theta_0)' \\ & + E\left\{a(\theta_0)\left[b(\theta_0)'/\sqrt{n} + c(\theta_0)'/n\right]\right\} + E\left\{\left[b(\theta_0)/\sqrt{n} + c(\theta_0)/n\right] a(\theta_0)'\right\} \\ & - E\left\{\tilde{a}(\theta_0)\left[\tilde{b}(\theta_0)'/\sqrt{n} + \tilde{c}(\theta_0)'/n\right]\right\} - E\left\{\left[\tilde{b}(\theta_0)/\sqrt{n} + \tilde{c}(\theta_0)/n\right]\tilde{a}(\theta_0)'\right\} \\ & + O_p(n^{-3/2}) \end{aligned}$$

The first term in the above expansion is the asymptotic variance. Since the MEL and MPP are first order equivalent it follows that $a(\theta_0) = \tilde{a}(\theta_0)$. The two estimators being $O_p(n^{-3/2})$ equivalent implies that $b(\theta_0) = \tilde{b}(\theta_0)$. Using these two basic results, the above

expansion for the difference becomes

$$\begin{aligned} & Var\left(\sqrt{n}(\hat{\theta}_{EL}-\theta_0)\right) - Var\left(\sqrt{n}(\hat{\theta}_{EL}-\theta_0)\right) \\ = \; & E\left\{a(\theta_0)\left[c(\theta_0)' - \tilde{c}(\theta_0)'\right]\right\}/n + E\left\{\left[c(\theta_0) - \tilde{c}(\theta_0)\right]a(\theta_0)'\right\}/n \\ & + O(n^{-3/2}) \end{aligned}$$

Both the MPP and the MEL are first order efficient. This implies that the two expectations above must be equal to zero. The asymptotic covariance of $[c(\theta_0)' - \tilde{c}(\theta_0)']$ and $a(\theta_0)$ must be equal to zero because otherwise there would be a matrix B such that $var(\hat{\theta}_{EL} + n^{-1/2}B(\hat{\theta}_{EL} - \hat{\theta}_{MPP})) < var(\theta_{EL})$ that contradict the first order efficiency of $\hat{\theta}$. Hence, $\hat{\theta}_{EL}$ and $\hat{\theta}_{MPP}$ have the same higher order variance, as required.

□

Proof of Theorem 8. The proof follows, *mutatis mutandis,* from Proposition 2 of Chernozhukov and Hong (2003).

□

References

Chamberlain, G. and Imbens, G. W. (2003). Nonparametric applications of bayesian inference. *Journal of Business and Economic Statistics*, 21(1):12–18.

Chernozhukov, V. and Hong, H. (2003). An mcmc approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.

Han, C. and Phillips, P. C. B. (2006). Gmm with many moment conditions. *Econometrica*, 74:147 – 192.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–54.

Ibragimov, I. A. and Has'minskii, R. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, Berlin.

Imbens, G. W. (1997). One-step estimators for over-identified generalized method of moments models. *Review of Economic Studies*, 64(3):359–83.

Jaynes, E. T. (1983). Prior probabilities. In Rosenkrantz, R. D., editor, *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, pages 114–130. D. Reidel Publishing Company, Boston, Massachusetts.

Kim, J. Y. (2002). Limited information likelihood and bayesian analysis. *Journal of Econometrics*, 107(1-2):175–93.

Kitamura, Y. (1997). Empirical likelihood methods with weakly dependent processes. *The Annals of Statistics*, 25:2084–2102.

Kitamura, Y., Tripathi, G., and Ahn, H. (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrics*, 72(6):1667 – 1714.

Kleibergen, F. (2005). Testing parameters in gmm without assuming that they are identified. *Econometrica*, 73:1103 – 1123.

Meyer, B. D., Viscusi, W. K., and Durbin, D. L. (1995). Workers' compensation and injury duration: Evidence from a natural experiment. *American Economic Review*, 85(3):322–40.

Monahan, J. F. and Boos, D. (1992). Proper likelihoods for Bayesian analysis. *Biometrika*, 79:271–278.

Newey, W. K. and McFadden, D. (1994). Estimation and inference in large samples. In Engle, R. and McFadden, D., editors, *Handbook of Econometrics*, pages 2113–2245, Amsterdam. North-Holland.

Newey, W. K. and Smith, R. J. (2004). Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica*, 72(1):219–55.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–49.

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, 57:1027–57.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, 22:300–325.

Ragusa, G. (2005). Alternatives to gmm: Properties of minimum divergence estimators. mimeo.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical methods*. Springer.

Rosenkrantz, R. D. (1977). *Inference, Method, and Decision: Towards a Bayesian Philosophy of Science*. Reidel, Boston.

Rubin, D. (1981). Bayesian bootstrap. *Annals of Statistics*, 9:130–34.

Schennach, S. C. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika*, 92(1):31–46.

Sims, C. A. (1996). Interview with christopher a. sims. *Journal of Business and Economic Statistics*, 20(4):448–49.

Stock, J. H. and Wright, J. H. (2000). Gmm with weak identification. *Econometrica*, 68(5):1055–96.