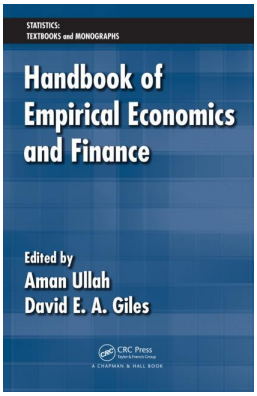

Handbook of Empirical Economics and Finance

Ullah Aman, E. A. Giles David

Large Deviations Theory and Econometric Information Recovery

Publication details
https://test.routledgehandbooks.com/doi/10.1201/b10440-7
Grendár Marian, Judge George
Published online on: 20 Dec 2010

6

Large Deviations Theory and Econometric Information Recovery

Marian Grendár and George Judge

CONTENTS

6.1 Estimation and Inference Base

Econometricians rarely have at their disposal enough information to formulate a model in terms of a parametric family of distributions. Consequently, traditional methods of parametric estimation and inference that are based either on a likelihood or on a posterior distribution are prone to committing

specification errors that lead to problems of inference. On the other hand, economic data are partial and incomplete and usually there is seldom a large enough data sample to rely on purely nonparametric methods.

Looking for a compromise, econometricians have turned to a traditional method of estimation and inference known as the *method of moments* (MM); cf., e.g., Mittelhammer, Judge, and Miller (2000). This formulation permits a researcher to specify only some moment properties/features of the data-sampling distribution F, with probability density function (pdf) $r(x)$, of a random variable $X \in \mathcal{R}^d$. This is accomplished through estimating functions $u(X; \theta) \in \mathcal{R}^J$ of parameter $\theta \in \Theta \subseteq \mathcal{R}^k$ (see Godambe and Kale 1991). The estimating functions are used to form a set $\Delta(\Theta) = \bigcup_{\theta \in \Theta} \Delta(\theta)$ of parametrized pdf's $\rho(x; \theta)$, defined through unbiased estimating equations (EE)

$$\Delta(\theta) = \left\{ \rho(x; \theta) : \int \rho(x; \theta) u(x; \theta) = 0 \right\}.$$

When $J = k$, there is usually a unique solution θ of the "just determined" set of EEs. Given a sample $X_1^n = X_1, \ldots, X_n$, the solution can be estimated by solving an empirical counterpart of the EEs: $\frac{1}{n} \sum_{i=1}^n u(x_i; \hat{\theta}_{\mathrm{MM}}) = 0$. The resulting estimator $\hat{\theta}_{\mathrm{MM}}$ is known as the *method of moments* estimator. Asymptotic distributional properties of the estimator are well known (Mittelhammer, Judge, and Miller 2000) and provide a basis for inference.

In econometric modeling, it often happens that there are more EEs than unknown parameters; $J > k$. A considerable amount of work has been devoted to extending the method of moments for this overdetermined case. As a result the Generalized Method of Moments (Hansen 1982) evolved (see also Hall 2005).

More recently (cf. Bickel et al. 1993; Mittelhammer, Judge, and Miller 2000; Owen 2001; among others), a new basis has emerged for regularizing the overdetermined EEs. This approach is based on minimization of a discrepancy, or divergence measure of a pdf ρ with respect to the true sampling distribution pdf r:

$$\phi(\rho \,\|\, r) = \int \phi\left(\frac{\rho(x)}{r(x)}\right) r(x) dx,$$

where ϕ is a convex function. If ρ is assumed to belong to model set $\Delta(\Theta)$ then the minimization problem

$$\hat{\rho}(\hat{\theta}) = \arg \inf_{\theta \in \Theta} \inf_{\rho(x;\, \theta) \in \Delta(\theta)} \phi(\rho \,\|\, r)$$

can be equivalently expressed in the convex dual form. Thanks to the convex duality, the optimal $\hat{\theta}$ can be obtained as

$$\hat{\theta} = \arg \inf_{\theta \in \Theta} \sup_{\gamma \in \mathcal{R}, \lambda \in \mathcal{R}^J} \gamma - \mathrm{E}[\phi^*(\gamma + \lambda' u(x; \theta))], \tag{6.1}$$

where ϕ^* is the convex conjugate of ϕ. In order to make Equation 6.1 operational, it is necessary to connect it with the sample data X_1^n. Indeed, in

Equation 6.1 the expectation is taken with respect to the true sampling distribution r. It is natural to replace the expectation by its sample-based estimate, and this leads to the empirical minimum divergence (EMD) estimator:

$$\hat{\theta}_{\text{EMD}} = \arg \inf_{\theta \in \Theta} \sup_{\gamma \in \mathcal{R}, \lambda \in \mathcal{R}^J} \gamma - \frac{1}{n} \sum_{i=1}^{n} [\phi^*(\gamma + \lambda' u(x_i; \theta))]. \tag{6.2}$$

Kitamura (2006) notes that the estimator (Equation 6.2) is the generalized minimum contrast estimator considered by Bickel et al. (1993).

There are two possible ways of using the parametric model $\Delta(\Theta)$, specified by EE.

1. One option is to use the EEs to define a feasible set of possible parametrized sampling distributions $q(x; \theta)$. In order to distinguish this way, the model set will be denoted $\Phi(\Theta)$. *The objective of information recovery is to select a representative sampling distribution from* $\Phi(\Theta)$. *This modeling strategy and associated problem will be referred to as a sampling distribution (SD) problem.*
2. Alternatively EEs can be used to form a set into which a parametrized empirical distribution should, in a researcher's view, belong. The sample X_1^n is used to estimate the sampling distribution r. In this case the model $\Delta(\Theta)$ will be denoted $\Pi(\Theta)$. *The objective of information recovery is the selection of a representative parametrized empirical distribution from* $\Pi(\Theta)$. *This modeling strategy and associated problem will be referred to as an empirical distribution (ED) problem.*

There are two choices of ϕ that are popular: (1) $\phi(x) = -\log(x)$ which leads[1] to the L-divergence (Grendár and Judge 2009a) $L(\rho \parallel r) = -\int r \log \rho$ of pdf ρ with respect to r, and (2) $\phi(x) = x \log(x)$ which leads to the I-divergence $I(\rho \parallel r) = \int \rho \log \frac{\rho}{r}$, which is also known as the Kullback Leibler divergence or the negative of relative entropy (Cover and Thomas 1991; Csiszár 1998). They both are members of the Cressie–Read (cf. Cressie and Read 1984; Cressie and Read 1988) (CR) parametric family of discrepancy measures

$$\phi(\rho \parallel r; \alpha) = \frac{1}{\alpha(\alpha + 1)} \int \left(\left(\frac{\rho(x)}{r(x)} \right)^{\alpha} - 1 \right) \rho(x) dx.$$

In the former case, the resulting EMD estimator is known as the empirical likelihood (EL) estimator (Mittelhammer, Judge, and Miller 2000; Owen 2001; Qin and Lawless 1994):

$$\hat{\theta}_{\text{EL}} = \arg \inf_{\theta \in \Theta} \sup_{\lambda \in \mathcal{R}^J} \frac{1}{n} \sum_{i=1}^{n} \log \left(1 - \lambda' u(x_i; \theta)\right). \tag{6.3}$$

[1] From the point of view of optimization of $\phi(\rho \parallel r)$ wrt ρ.

In the latter case the empirical maximum maximum entropy (EMME) or the maximum entropy empirical likelihood results (Back and Brown 1990; Imbens, Spady, and Johnson 1998; Kitamura and Stutzer 1997; Mittelhammer, Judge, and Miller 2000; Owen 2001):

$$\hat{\theta}_{\text{EMME}} = \arg\sup_{\theta\in\Theta}\inf_{\lambda\in\mathcal{R}^J}\frac{1}{n}\sum_{i=1}^{n}\exp(-\lambda' u(x_i;\theta)).$$

Asymptotic distributional properties of both estimators are known (cf., e.g., Mittelhammer, Judge, and Miller 2000; Owen 2001) and an inferential basis follows. Other members of the CR class of discrepancies appear in the literature; cf., e.g., Schennach (2007). It is also possible to select an optimal EMD estimator from the CR class, where optimality may be suitably defined by minimum mean squared error criterion or some other loss function (Judge and Mittelhammer 2004; Mittelhammer and Judge 2005). A survey of the known small and large sample properties of EMD estimators can be found in, e.g., Schennach (2007) and Grendár and Judge (2008); see also Owen (2001) and Mittelhammer, Judge, and Miller (2000).

In practice, an econometrician usually does not have enough information to guarantee that the model set (either Φ, or Π) contains the true data-sampling distribution r. Given that most econometric-statistical models are misspecified, it is of basic interest to know which of the methods of information recovery is consistent in the misspecified case. This is the place where the large deviations (LD) theory (cf., e.g., Dembo and Zeitouni 1998; Cover and Thomas 1991) is of great use, as it permits us to find out, in both the ED and SD settings, the methods that are consistent under misspecification. This way LD provides a guidance in information recovery, as it shows which methods are ruled out and in what sense.

6.1.1 Purpose-Objectives of This Chapter

In the context of the above estimation and inference base this chapter provides a nontechnical[2] introduction to LD theory with a focus on the implications of some of the key LD theorems for information recovery in *Econometrics* and *Statistics*. LD theory is a subfield of probability theory where, informally, the typical concern is about the asymptotic behavior, on a logarithmic scale, of the probability of a given event. In more technical words, LD theory studies the exponential decay of the probability of an event. For example, consider the event that an empirical measure belongs to a specified set. The rate of exponential decay of the probability of this event is determined by an extremal value of a certain quantity, called the rate function, over the set; cf. the Sanov theorem (ST), Subsection 6.2.1. This permits one to estimate the probability of this event with precision of the first order in the exponent. Even more importantly, ST leads to the conditional law of large numbers (CLLN), which says,

[2] Theorems are stated without proof, but references to the literature where the proofs can be found are provided. Theorems are intentionally not stated at greatest possible generality.

given that the event has occurred, only empirical measures arbitrarily close to those that minimize the rate function, can occur as the sample size goes to infinity.

Although LD theory also studies other types of events, we concentrate on ST as a means for establishing CLLN. CLLN has profound implications for the relative entropy maximization (REM/MaxEnt) – since, in the i.i.d. case, the rate function is just the Kullback Leibler information divergence. These implications carry over in a parametric context to EEs, where they provide a probabilistic justification (cf. Kitamura and Stutzer 2002) to empirical MaxMaxEnt estimation method (or exponential tilt estimator).

The lack of a comparable probabilistic justification for the empirical likelihood approach has motivated a study of LD theorems for data-sampling distributions: Bayesian Sanov theorem (BST) and its corollary: Bayesian law of large numbers (BLLN). The other objective of this chapter is to expose, in general, the relatively new LD theorems for sampling distributions (also known as Bayesian LD theorems) and their implications. Bayesian LLN provides a probabilistic justification to the maximum nonparametric likelihood (MNPL) method, which carries over in a parametric context, where it justifies the estimation method known as empirical likelihood (EL). The BLLN also shows that from an estimation point of view MNPL as well as EL can be seen as asymptotic instances of Bayesian maximum a posteriori probability (MAP) method, in nonparametric and parametric context, respectively.

6.1.2 Organization of the Chapter

The chapter is divided into two large sections (Sections 6.2 and 6.4), which are connected by an Intermezzo (Section 6.3). Section 6.2 is devoted to explaining key LD theorems for empirical measures, culminating with CLLN and its implications for ED selection problems. In particular, in Subsection 6.2.1, the Sanov theorem is stated, explained, and illustrated with an example. Then we demonstrate that the law of large numbers (LLN) is a direct consequence of ST that implies how the simplest problem of selection of ED should be solved. This is intended to facilitate understanding of how the CLLN provides a probabilistic justification to relative entropy maximization and maximum probability methods, in context of the general ED problem. Next we step by step extend the ED problem into parametric and then to empirical parametric ED problems. CLLN is used to determine regularization methods. Not surprisingly, the methods are MaxMaxEnt, empirical MaxMaxEnt (also known as maximum entropy empirical likelihood or exponential tilt), respectively. Finally, the continuous case of the empirical parametric ED problem is addressed. Intermezzo (Section 6.3) summarizes implications of CLLN for the parametric ED selection problems and prepares a transition to the opposite SD selection problems. Next, Section 6.4 presents several LD theorems for sampling distributions (including the Bayesian Sanov theorem and the Bayesian law of large numbers), and discusses implications of BLLN for the SD problem in its basic as well as in its parametric forms. A summary is followed by some literature notes.

6.2 Large Deviations for Empirical Distributions

In order to discuss the ST CLLN, it is necessary to introduce some basic terminology; cf. Csiszár (1998).

Let $\mathcal{P}(\mathcal{X})$ be a set of all probability mass functions on the finite, m-element set $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$. The support of $p \in \mathcal{P}(\mathcal{X})$ is a set $S(p) = \{x : p(x) > 0\}$.

Let $x_1, x_2, \ldots, x_n$ be a random sample from a pmf $q \in \mathcal{P}(\mathcal{X})$. Let ν^n denote the empirical measure induced by a random sample of length n. Formally, the empirical measure $\nu^n = [n_1, n_2, \ldots, n_m]/n$, where n_i is the number of occurrences of ith element of $\mathcal{X}$ in the random sample. When there is a need to stress the size n of the random sample that induces the empirical measure, we will speak about the n-empirical measure. Finally, note that there are $\Gamma(\nu^n) = n!(\prod_{i=1}^m n_i!)^{-1}$ different random samples of length n that induce the same empirical measure ν^n.

As previously mentioned, we are interested in the event that the random sample drawn from a fixed pmf q induces the empirical measure ν^n from a set $\Pi \subseteq \mathcal{P}(\mathcal{X})$. The probability of this event is therefore

$$\pi(\nu^n \in \Pi; q) = \sum_{\nu^n \in \Pi} \pi(\nu^n; q),$$

where

$$\pi(\nu^n; q) = \Gamma(\nu^n) \prod_{i=1}^m q_i^{n_i}.$$

The Large Deviations (LD) rate function of probability of this event is the information divergence. The information divergence (I-divergence, $\pm$-relative entropy, Kullback Leibler distance, etc.) $I(p \parallel q)$ of $p \in \mathcal{P}(\mathcal{X})$ with respect to $q \in \mathcal{P}(\mathcal{X})$ is

$$I(p \parallel q) = \sum_{i=1}^m p_i \log \frac{p_i}{q_i},$$

where $0 \log 0 = 0$ and $\log b/0 = +\infty$, by convention. The information projection $\hat{p}$ of q on Π is

$$\hat{p} = \arg \inf_{p \in \Pi} I(p \parallel q).$$

The I-divergence at an I-projection of q on Π is denoted $I(\Pi \parallel q)$. Finally, recall that $I(p \parallel q) \geq 0$, where $I(p \parallel q) = 0$ iff $p = q$.

Topological qualifiers, e.g. openness, will be used with respect to the topology induced on $\mathcal{P}(\mathcal{X})$ by the standard topology on $\mathcal{R}^m$.

6.2.1 Sanov Theorem

The Sanov theorem, which is the basic LD result on the asymptotic behavior of the probability that an n-empirical measure from a specified set Π occurs, may be stated as:

Sanov Theorem *Let Π be an open set and let $S(q) = \mathcal{X}$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log \pi(\nu^n \in \Pi; q) = -I(\Pi \| q).$$

Phrased informally, ST tells us that the rate of the exponential convergence of the probability $\pi(\nu^n \in \Pi; q)$ toward zero, is determined by the information divergence at (any of) the I-projection(s) of q on Π. The other probability mass functions (pmf's) do not influence the rate of convergence. Obviously, the greater the value $I(\Pi \| q)$, the faster the convergence of probability $\pi(\nu^n \in \Pi; q)$ to zero, as $n \to \infty$. The Sanov theorem thus permits us to speak about and measure how rare a set Π is with respect to a sampling distribution.

We use Example 6.1 to illustrate ST.

Example 6.1

Let $\mathcal{X} = [1, 2, 3, 4]$. Let q be the uniform pmf. Let $\Pi = \{p : \sum_{i=1}^{4} p_i x_i = 3.0\}$. The Table 6.1 illustrates the convergence of the normalized log-probability $\frac{1}{n} \log \pi(\nu^n \in \Pi; q)$ to $-I(\Pi \| q)$, as $n \to \infty$.

The probability that a 1000-empirical measure with a mean of 3.0 was drawn from q is $\pi(\nu^{1000} \in \Pi; q) = 4.1433e-47$. An approximate estimate of the probability can be obtained by means of ST, as $\pi(\nu^{1000} \in \Pi; q) \approx \exp(-1000\, I(\Pi \| q)) = 3.4121e-45$. Note that the approximation is comparable to the exact probability.

6.2.2 Law of Large Numbers

Since the probability $\pi(\nu^n \in A; q)$ goes to zero for any $A \subset \mathcal{P}(\mathcal{X})$ such that $I(A \| q) > 0$, the use of ST leads to the following law of large numbers (LLN):

Law of Large Numbers *Let $B(\hat{p}, \epsilon)$ be the closed ϵ-ball defined by the total variation metric and centered at the I-projection $\hat{p} \equiv q$ of q on $\mathcal{P}(\mathcal{X})$. Then,*

$$\lim_{n\to\infty} \pi(\nu^n \in B(\hat{p}, \epsilon); q) = 1.$$

TABLE 6.1

Convergence of the Normalized Log-Probability to the Negative of Minimum Value of the I-Divergence

n	$1/n \log \pi(\nu^n \in \Pi; q)$
10	−0.3152913
100	−0.1350160
1000	−0.1068000
$-I(\Pi \| q)$	−0.1023890

Thus, the information divergence that, through its infimal value, gave the rate of exponential decay also provides, through the "point" of its infimum, the pmf around which n-empirical measures concentrate. This LLN may be interpreted as saying that, asymptotically, the only possible empirical measures are those that are arbitrary close to the I-projection $\hat{p}$ of q on $\mathcal{P}(\mathcal{X})$, i.e., in this case, to the data-sampling distribution q.

Next, we consider the example showing how LLN implies that the simplest problem of selection of ED has to be solved using the REM/MaxEnt method. The ED problem concerns selection of empirical measure from Π, when the information-quadruple $(\mathcal{X}, q, n, \Pi)$ and nothing else is available. The next Example illustrates the ED problem, which is simplest in the sense that Π is identical with the entire $\mathcal{P}(\mathcal{X})$.

Example 6.2
Let $\mathcal{X} = \{1, 2, 3, 4\}$ *and let* $q = [0.1, 0.4, 0.2, 0.3]$. *Let a random sample of size* $n = 10^9$ *be drawn from* q. *The sample is not available to us. We are told only that the sample mean is somewhere in the interval* $[1, 4]$; *i.e.,* $\Pi = \{p : \sum p_i x_i \in [1, 4]\}$. *Given the information-quadruple* $(\mathcal{X}, q, n, \Pi)$ *we are asked to select an n-empirical measure from* Π.

Since any 10^9-empirical measure fits the given interval of mean values, there are $N = \binom{n+m-1}{m-1}$ empirical measures from which we are asked to make a choice. The problem that Example 6.2 asks us to solve is in the form of an under-determined, ill-posed inverse problem.

The information that the unknown n-empirical measure was drawn from q is crucial for comprehending that the ill-posed problem has a simple solution implied by LLN. Though there are N n-empirical measures in Π, LLN implies that only those n-empirical measures that are close to the I-projection $\hat{p} \equiv q$ of the sampling distribution q on $\mathcal{P}(\mathcal{X})$, i.e., close to q, are possible. Hence, LLN regularizes the ill-posed inverse problem. Since the relative entropy maximization method (REM/MaxEnt) selects just the I-projection, REM must be used to solve this problem.

Let us conclude by noting that a consistency requirement would imply that the same method should also be used for "small" n. Thus, if the sample size were $n = 10$ instead of 10^9, as in Example 6.2, consistency would imply that one should select the I-projection of q on the set $\mathcal{P}_{10}(\mathcal{X})$ of all possible 10-empirical measures. Less stringently viewed, LLN implies that any method that asymptotically becomes identical to REM can be used for solving this instance of the ED problem.

6.2.3 CLLN, Maximum Entropy, and Maximum Probability

To demonstrate that LLN is a special case of the CLLNs, it is instructive to express the claim of LLN in the following form:

$$\lim_{n\to\infty} \pi(\nu^n \in B(\hat{p}, \epsilon) \mid \nu^n \in \mathcal{P}(\mathcal{X});\ q) = 1.$$

Compare this to the result of CLLN:

Conditional Law of Large Numbers *Let* Π *be a convex, closed set that does not contain* q*. Let* $B(\hat{p}, \epsilon)$ *be a closed* ϵ*-ball defined by the total variation metric and centered at the* I*-projection* $\hat{p}$ *of* q *on* Π*. Then,*

$$\lim_{n\to\infty} \pi(\nu^n \in B(\hat{p}, \epsilon) \mid \nu^n \in \Pi; q) = 1.$$

Thus, one can see that CLLN generalizes LLN. Interpretation of CLLN is similar to that of LLN except that the conditioning set Π is no longer the entire $\mathcal{P}(\mathcal{X})$ but a convex, closed subset that does not contain q; hence the model Π is misspecified. In other words, given the conditioning set Π, CLLN demonstrates that empirical measures asymptotically conditionally concentrate on the I-projection of q on Π, provided that the set satisfies certain technical requirements.

CLLN follows directly from the Sanov theorem. The conditional probability in question is

$$\pi(\nu^n \in B(p, \epsilon) \mid \nu^n \in \Pi;\ q) = \frac{\pi(\nu^n \in B(p, \epsilon); q)}{\pi(\nu^n \in \Pi; q)}.$$

Hence, by ST, if $B(p, \epsilon)$ is such that $\hat{p} \notin B(p, \epsilon)$ then

$$\frac{1}{n} \log \pi(\nu^n \in B(p, \epsilon) \mid \nu^n \in \Pi;\ q) \to -\underbrace{(\, I(B(p, \epsilon) \,\|\, q) - I(\Pi \,\|\, q)\,)}_{>\,0}$$

and consequently $\pi(\nu^n \in B(p, \epsilon) \mid \nu^n \in \Pi;\ q) \to 0$. Since, by assumption, there is a unique I-projection of q on Π, the conditional probability concentrates on it.

In Subsection 6.2.2, LLN was invoked to solve Example 6.2, a simple instance of the ED problem. The following extension of Example 6.2 provides a more general instance of the problem in the sense that Π is now a subset of $\mathcal{P}(\mathcal{X})$.

Example 6.3
Let $\mathcal{X} = \{1, 2, 3, 4\}$ *and let* $q = [0.1, 0.4, 0.2, 0.3]$*. Let a random sample of size* $n = 10^9$ *be drawn from* q*. The sample is not available to us. What we are told, only, is that the sample mean is* 3.0*. Note that it is different from the expected value of* X *under the pmf* q*,* 2.7*; i.e., the model is misspecified. Hence, the feasible set to which n-types belong is* $\Pi = \{p : \sum_{i=1}^{4} p_i x_i = 3.0\}$*. Given the information-quadruple* $(\mathcal{X}, q, n, \Pi)$*, how should one go about selecting an n-empirical measure from* Π*?*

The same discussion as that following Example 6.2 applies to this example except that now it is CLLN instead of LLN that regularizes the ill-posed ED problem.

Example 6.3 (cont'd)
CLLN dictates that one selects an empirical measure close to the I*-projection* $\hat{p} = [0.057,\ 0.310,\ 0.209,\ 0.424]$ *of* q *on* Π*, which is now different from* q*. The* I*-projection can be obtained in the standard way of solving the constrained relative*

entropy maximization task (cf., e.g., Golan, Judge, and Miller 1996): $\hat{p} = \arg\max_{p\in\Pi} -\sum p_i \log \frac{p_i}{q_i}$, *where* $\Pi = \{p : \sum_{i=1}^{4} p_i x_i = 3.0, \sum_{i=1}^{4} p_i = 1\}$.

The following is an example of the ED problem with an economic relevance.

Example 6.4

(Cox, Daniell, and Nicole 1998) studied the UK National Lottery, where every week, 6 *numbers from* 49 *are drawn, at random. The Lottery makes available the following information about a draw: the winning ticket* s*, the total number* n *of sold tickets, number of winners* n_r *in each category (i.e., matched* r*-tuple),* $r = 3, 4, 5, 6$*. The info is available for* W *weeks. The authors assume that the distribution* $q(t)$ *of tickets is uniform.*

Given the above information, the objective is to select a representative empirical distribution of tickets. This is an instance of the problem of ED selection, where the feasible set of empirical pmf's is formed by the available information as follows: for each draw (week) w*, winning ticket* s *and category* $r = 3, 4, 5$*,*

$$n_r(w) = \sum_t \delta_r(t, s)n(t),$$

where $n(t)$ *is the unknown number of people who bought the ticket* t *and* $\delta_r(t, s) = 1$*, if* t *and* s *have common just* r *numbers,* 0 *otherwise. There are* $3W$ *constraints.*

The authors used information from $W = 113$ *weeks and regularized the* Π*-problem by relative entropy maximization method. In the Table 1.1 (of Cox, Daniell, and Nicole 1998) one can find, for instance, that ticket with numbers* 26 34 44 46 47 49 *has estimated* $n(t) = 0.41$ *while the ticket* 7 17 23 32 40 42 *has estimated* $n(t) = 45.62$*; on an appropriate scale.*

This work triggered a new economic interest in lotteries; cf., for instance, Farrell et al. (2000).

In conclusion, LD theory for empirical distributions, through the CLLNs provides a probabilistic justification for using REM to solve the ED selection problem. Alternatively, any method of solving the ED problem that asymptotically does not behave like REM, violates CLLN. For instance, using the maximum Tsallis entropy method (maxTent) to solve the Π-problem would go against CLLN. However, using the Maximum Probability method (MaxProb) (Boltzmann 1877; Vincze 1972; Grendár and Grendár 2001; Niven 2007) satisfies CLLN since it asymptotically turns into REM (cf. Grendár and Grendár 2001, 2004). The MaxProb method suggests that one may solve the ED problem by selecting the μ-projection of q on Π, i.e., the n-empirical measure

$$\nu^n_{\text{MaxProb}} = \arg\sup_{\nu^n\in\Pi} \pi(\nu^n; q).$$

It is worth noting that the convergence of the most probable empirical measure(s), obtained with MaxProb, to the distribution that maximizes relative entropy, obtained with REM/MaxEnt, provides another, deeper reading of CLLN, namely that the empirical measures conditionally concentrate on the asymptotically most probable empirical measure.

Finally, a word of caution: The information-divergence minimization method (REM/MaxEnt) is also used to regularize problems like spectrum estimation or recovering of X-ray attenuation functions or optical images (cf. Jones and Byrne 1990), that cannot be cast into the form of the Π-problem. In such cases, the LD justification used for REM cannot be invoked, and one has to rely on other arguments; cf. Jones and Byrne (1990) and Csiszár (1996).

6.2.4 Parametric ED Problem and Maximum Maximum Entropy Method

The feasible set Π of the ED problem can be, in general, defined by means of J moment-consistency constraints $\Pi = \{p : \sum_{i=1}^{m} p_i u_j(x_i) = a_j,\ j = 1, 2, \ldots, J\}$, where $u_j(\cdot)$ is a real-valued function of X called the u-moment and $a \in \mathcal{R}^J$ is given. In this case, the I-projection of q on Π is easy to find, as it belongs to the exponential family of distributions

$$\mathcal{E}(X, \lambda, u) = k(\lambda) q(X) \exp\left(-\sum_{j=1}^{J} \lambda_j u_j(X)\right),$$

where $k(\lambda) = (\sum_{i=1}^{m} q(x_i) \exp(-\sum_{j=1}^{J} \lambda_j u_j(x_i)))^{-1}$ is the normalizing constant. Example 6.3 presents a simple Π with a single u-moment of the form $u(X) = X$.

The u-moment function $u(X)$ can be viewed as a special case of a general, parametric $u(X, \theta)$-moment function, where θ is a parameter, $\theta \in \Theta \subseteq \mathcal{R}^k$. The parametric u-moments define the parametric feasible set $\Pi(\Theta) = \bigcup_{\theta \in \Theta} \Pi(\theta)$, where

$$\Pi(\theta) = \left\{ p(\cdot; \theta) : \sum_{i=1}^{m} p(x_i; \theta) u_j(x_i, \theta) = 0,\ j = 1, 2, \ldots, J \right\}. \tag{6.4}$$

Example 6.5 illustrates the extension of the ED problem to the parametric ED problem.

Example 6.5

Let $\mathcal{X} = \{1, 2, 3, 4\}$ *and let* $q = [0.1, 0.4, 0.2, 0.3]$. *Let a random sample of size* $n = 10^9$ *be drawn from* q. *The sample is unavailable to us. What we are told, only, is that the sample mean is in the interval* $[3.0, 4.0]$. *Here,* $u(X, \theta) = X - \theta$, *where* $\theta \in [3.0, 4.0]$; *so that* $\Pi(\theta) = \{p(\cdot; \theta) : \sum p(x_i; \theta)(x_i - \theta) = 0\}$ *and* $\theta \in \Theta = [3.0, 4.0]$. *Note that the interval does not contain the expected value of* X *with respect to* q, $E_q X = 2.7$; *i.e., the model* $\Pi(\Theta)$ *is misspecified. The objective is to select a parametrized n-empirical measure* $\nu^n(\theta)$ *from* $\Pi(\Theta)$, *given the available information.*

Let us link the parametric Π-problem to the estimating equations (EE) approach to estimation. The general, parametric $u(X, \theta)$-moment function is, in this context, commonly known as an estimating function. Unbiased estimating functions are the most commonly considered estimating functions in

Econometrics. Thus, the parametric ED problem becomes a problem of estimating the unknown value of θ. In other words, given a scheme for selecting $\hat{p}(\cdot; \hat{\theta})$, we are more interested in $\hat{\theta}$ than in the corresponding $\hat{p}(\cdot; \hat{\theta})$. It is clear that the selected $\hat{\theta}$ will depend on q.

LD theory provides a clue as to how one should solve the parametric ED problem. If $\Pi(\Theta)$ is a convex, closed set that does not contain q (i.e., the model is misspecified), CLLN can be invoked to claim that such a parametric Π-problem should be solved by selecting the I-projection $\hat{p}(\cdot; \hat{\theta})$ of the sampling distribution q on $\Pi(\Theta)$, i.e.,

$$\hat{p}(\cdot; \theta) = \arg \inf_{p(\cdot;\theta)\in\Pi(\theta)} I(p(\cdot; \theta) \,\|\, q)$$

with $\theta = \hat{\theta}_{\mathrm{MME}}$, where

$$\hat{\theta}_{\mathrm{MME}} = \arg \inf_{\theta\in\Theta} I(\hat{p}(\cdot; \theta) \,\|\, q).$$

Because of the double maximization of the entropy, we will call the method associated with this prescription Maximum Maximum Entropy method (MaxMaxEnt). If $\Pi(\Theta)$ is defined by the (Equation 6.4) the estimator $\hat{\theta}_{\mathrm{MME}}$ can be expressed as

$$\hat{\theta}_{\mathrm{MME}} = \arg \sup_{\theta\in\Theta} \inf_{\lambda\in\mathcal{R}^l} \log \sum_{i=1}^{m} q(x_i; \theta) \exp\left(-\lambda' u(x_i; \theta)\right).$$

Example 6.5 (cont'd)
Note that MaxMaxEnt when applied to Example 6.5 selects the same pmf as did MaxEnt in Example 6.3. Indeed, since the information divergence is a convex function in the first argument, the minimum of the information divergence over $\theta \in [3.0, 4.0]$ *is attained for* $\theta = 3.0$. *Phrased in EE terms, the MaxMaxEnt estimator of the unknown true value of* θ, *based on the available information is* $\hat{\theta}_{MME} = 3.0$.

6.2.5 Empirical ED Problem

The setting of the Π-problem is idealized. In practice, the data-sampling distribution q is rarely known. Let us continue assuming that the other components of the information-quadruple that constitute the ED problem, i.e., Π, n (the size of the sample that is unavailable to us) and $\mathcal{X}$, are known to us. To make the setup and problem more realistic, imagine that we draw a random sample $X_1^N = X_1, X_2, \ldots, X_N$ of size N from the true data-sampling distribution q. Let the sample induce the N-empirical measure ν^N. When q in the information-quadruple is replaced by ν^N, we speak about the empirical ED problem. CLLN implies that the empirical ED problem should be solved by any method whose choice becomes asymptotically (i.e., as $n \to \infty$) identical with the I-projection of ν^N on Π.

6.2.6 Empirical Parametric ED Problem and Empirical MaxMaxEnt

The discussion of Subsection 6.2.5 extends directly to the empirical parametric ED problem, which CLLN implies should be solved by selecting

$$\hat{p}(\cdot;\theta) = \arg \inf_{p(\cdot;\theta)\in\Pi(\theta)} I(p(\cdot;\theta) \parallel \nu^N)$$

with $\theta = \hat{\theta}_{\text{EMME}}$, where

$$\hat{\theta}_{\text{EMME}} = \arg \inf_{\theta\in\Theta} I(\hat{p}(\cdot;\theta) \parallel \nu^N).$$

The estimator $\hat{\theta}_{\text{EMME}}$ is known in *Econometrics* under various names such as maximum entropy empirical likelihood and exponential tilt. We call it the empirical MaxMaxEnt estimator (EMME). Note that thanks to the convex duality, the estimator $\hat{\theta}_{\text{EMME}}$ can equivalently be obtained as

$$\hat{\theta}_{\text{EMME}} = \arg \sup_{\theta\in\Theta} \inf_{\lambda\in\mathcal{R}^l} \log \sum_{i=1}^{m} \nu^N(x_i;\theta) \exp\left(-\lambda' u(x_i;\theta)\right). \tag{6.5}$$

Example 6.6 illustrates the extension of the parametric ED problem (cf. Example 6.5) to the empirical parametric ED problem.

Example 6.6

Let $\mathcal{X} = \{1, 2, 3, 4\}$. Let a random sample of size $N = 100$ from data-sampling distribution q induces N-type $\nu^N = [7\ 42\ 24\ 27]/100$. Let in addition a random sample of size $n = 10^9$ be drawn from q, but it remains unavailable to us. We are told only that the sample mean is in the interval $[3.0, 4.0]$. Thus $\Pi(\theta) = \{p(\cdot;\theta) : \sum_{i=1}^{4} p(x_i;\theta)(x_i - \theta) = 0\}$ and $\theta \in \Theta = [3.0, 4.0]$. The objective is to select an n-empirical measure from $\Pi(\Theta)$, given the available information.

CLLN dictates that we solve the problem by EMME. Since n is very large, we can without much harm ignore rational nature of n-types (i.e., $\nu^n(\cdot;\theta) \in \mathcal{Q}^m$) and seek the solution among pmf's $p(\cdot;\theta) \in \mathcal{R}^m$. CLLN suggests the selection of $\hat{p}(\hat{\theta}_{EMME})$. Since the average $\sum_{i=1}^{4} \nu_i^N x_i = 2.71$, is outside of the interval $[3.0, 4.0]$, convexity of the information divergence implies that $\hat{\theta}_{EMME} = 3.0$, i.e., the lower bound of the interval.

Kitamura and Stutzer (2002) were the first to recognize that LD theory, through CLLN, can provide justification for the use of the EMME estimator. The CLLNs demonstrate that selection of I-projection is a consistent method, which in the case of a parametric, possibly misspecified model $\Pi(\Theta)$, establishes consistency under misspecification of the EMME estimator.

Let us note that ST and CLLN have been extended also to the case of continuous random variables; cf. Csiszár (1984); this extension is outside the scope of this chapter. However, we note that the theorems, as well as Gibbs conditioning principle (cf. Dembo and Zeitouni 1998) and Notes on literature),

when applied to the parametric setting, single out

$$\hat{\theta}_{\text{EMME}} = \arg\sup_{\theta\in\Theta}\inf_{\lambda\in\mathcal{R}^J}\frac{1}{N}\sum_{l=1}^{N}\exp\left(-\lambda' u(x_l;\theta)\right) \tag{6.6}$$

as an estimator that is consistent under misspecification. The estimator is the continuous-case form of Empirical MaxMaxEnt estimator. Note that the above definition (Equation 6.6) of the EMME reduces to Equation 6.5, when X is a discrete random variable. In conclusion it is worth stressing that in ED-setting the EMD estimators from the CR class (cf. Section 6.1) other than EMME are not consistent, if the model is not correctly specified.

A setup considered by Qin and Lawless (1994) (see also Grendár and Judge 2009b) serves for a simple illustration of the empirical parametric ED problem for a continuous random variable.

Example 6.7
Let there be a random sample from a (unknown to us) distribution $f_X(x)$ on $\mathcal{X} = \mathcal{R}$. We assume that the data were sampled from a distribution that belongs to the following class of distributions (Qin and Lawless 1994): $\Pi(\theta) = \{p(x;\theta) : \int_{\mathcal{R}} p(x;\theta)(x-\theta)\,dx = 0, \int_{\mathcal{R}} p(x;\theta)(x^2-(2\theta^2+1))\,dx = 0, p(x;\theta)\in\mathcal{P}(\mathcal{R})\}$, *and* $\theta\in\Theta = \mathcal{R}$. *However, the true sampling distribution need not belong to the model* $\Pi(\Theta)$. *The objective is to select a* $p(\theta)$ *from* $\Pi(\Theta)$. *The large deviations theorems mentioned above single out* $\hat{p}(\hat{\theta}_{EMME})$, *which can be obtained by means of the nested optimization* (*Equation 6.6*).

For further discussions and application of EMME to asset pricing estimation, see Kitamura and Stutzer (2002).

6.3 Intermezzo

Since we are about to leave the area of LD for empirical measures for the, in a sense, opposite area of LD for data-sampling distributions, let us pause and recapitulate the important points of the above discussions.

The Sanov theorem, which is the basic result of LD for empirical measures, states that the rate of exponential convergence of probability $\pi(\nu^n\in\Pi; q)$ is determined by the infimal value of information divergence (Kullback-Leibler divergence) $I(p\,\|\,q)$ over $p\in\Pi$. Though seemingly a very technical result, ST has fundamental consequences, as it directly leads to the law of large numbers and, more importantly, to its extension, the CLLNs (also known as the conditional limit theorem). Phrased in the form implied by Sanov theorem, LLN says that the empirical measure asymptotically concentrates on the I-projection $\hat{p}\equiv q$ of the data-sampling q on $\Pi\equiv\mathcal{P}(\mathcal{X})$. When applying LLN, the feasible set of empirical measures Π is the entire $\mathcal{P}(\mathcal{X})$. It is of interest to know the point of concentration of empirical measures when Π is a subset

of $\mathcal{P}(\mathcal{X})$. Provided that Π is a convex, closed subset of $\mathcal{P}(\mathcal{X})$, this guarantees that the I-projection is unique. Consequently, CLLN shows that the empirical measure asymptotically conditionally concentrates around the I-projection $\hat{p}$ of the data-sampling distribution of q on Π. *Thus, the CLLNs regularizes the ill-posed problem of ED selection. In other words, it provides a firm probabilistic justification for the application of the relative entropy maximization method in solving the ED problem*. We have gradually considered more complex forms of the problem, recalled the associated conditional laws of large numbers, and showed how CLLN also provides a probabilistic justification for the empirical MaxMaxEnt method (EMME). It is also worth recalling that any method that fails to behave like EMME asymptotically would violate CLLN if it were used to obtain a solution to the empirical parametric ED problem.

6.4 Large Deviations for Sampling Distributions

Now, we turn to a corpus of "opposite" LD theorems that involves LD theorems for data-sampling distributions, which assume a Bayesian setting. First, the Bayesian Sanov theorem (BST) will be presented. We will then demonstrate how this leads to the Bayesian law of large numbers (BLLN). These LD theorems for sampling distributions will be linked to the problem of selecting a sampling distribution (SD problem, for short). We then demonstrate that if the sample size n is sufficiently large the problem should be solved with the maximum nonparametric likelihood (MNPL) method. As with the problem of empirical distribution (ED) selection, requiring consistency implies that the SD problem should be solved with a method that asymptotically behaves like MNPL. The Bayesian LLN implies that, for finite n, there are at least two such methods, MNPL itself and maximum a posteriori probability. Next, it will be demonstrated that the Bayesian LLN leads to solving the parametric SD problem with the empirical likelihood method when n is sufficiently large.

6.4.1 Bayesian Sanov Theorem

In a Bayesian context assume that we put a strictly positive prior probability mass function $\pi(q)$ on a countable[3] set $\Phi \subset \mathcal{P}(\mathcal{X})$ of probability mass functions (sampling distributions) q. Let r be the "true" data-sampling distribution, and let X_1^n denote a random sample of size n drawn from r. Provided that $r \in \Phi$, the posterior distribution

$$\pi(q \in Q \mid X_1^n = x_1^n; r) = \frac{\sum_Q \pi(q) \prod_{i=1}^n q(x_i)}{\sum_\Phi \pi(q) \prod_{i=1}^n q(x_i)}$$

[3] We restrict presentation to this case, in order to not obscure it by technicalities; cf. Grendár and Judge (2009a) for Bayesian LD theorems in a more general case and more complete discussions.

is expected to concentrate in a neighborhood of the true data-sampling distribution r as n grows to infinity. Bayesian nonparametric consistency considerations focus on exploration of conditions under which it indeed happens; for entries into the literature we recommend Ghosh and Ramamoorthi (2003); Ghosal, Ghosh, and Ramamoorthi (1999); Walker (2004); and Walker, Lijoi, and Prünster (2004), among others. Ghosal, Ghosh, and Ramamoorthi (1999) define consistency of a sequence of posteriors with respect to a metric or discrepancy measure d as follows: The sequence $\{\pi(\cdot \mid X_1^n; r), n \geq 1\}$ is said to be d-consistent at r, if there exists a $\Omega_0 \subset \mathcal{R}^\infty$ with $r(\Omega_0) = 1$ such that for $\omega \in \Omega_0$, for every neighborhood U of r, $\pi(U \mid X^n; r) \to 1$ as n goes to infinity. If a posterior is d-consistent for any $r \in \Phi$, then it is said to be d-consistent. Weak consistency and Hellinger consistency are usually studied in the literature.

Large deviations techniques can be used to study Bayesian nonparametric consistency. The Bayesian Sanov theorem identifies the rate function of the exponential decay. This in turn identifies the sampling distributions on which the posterior concentrates, as those distributions that minimize the rate function. In the i.i.d. case the rate function can be expressed in terms of the L-divergence. The L-divergence (Grendár and Judge 2009a) $L(q \parallel p)$ of $q \in \mathcal{P}(\mathcal{X})$ with respect to $p \in \mathcal{P}(\mathcal{X})$ is defined as

$$L(q \parallel p) = -\sum_{i=1}^{m} p_i \log q_i.$$

The L-projection $\hat{q}$ of p on $A \subseteq \mathcal{P}(\mathcal{X})$ is

$$\hat{q} = \arg\inf_{q \in A} L(q \parallel p).$$

The value of L-divergence at an L-projection of p on A is denoted by $L(A \parallel p)$. Finally, let us stress that in the discussion that follows, r need not be from Φ; i.e., we are interested in Bayesian nonparametric consistency under misspecification.

In this context the Bayesian Sanov theorem (BST) provides the rate of the exponential decay of the posterior probability.

Bayesian Sanov Theorem *Let $Q \subset \Phi$. As $n \to \infty$,*

$$\frac{1}{n} \log \pi(q \in Q \mid x_1^n; r) \to -\{L(Q \parallel r) - L(\Phi \parallel r)\}, \quad \textit{a.s. } r^\infty.$$

In effect BST demonstrates that the posterior probability $\pi(q \in Q \mid x_1^n; r)$ decays exponentially fast (almost surely), with the decay rate specified by the difference in the two extremal L-divergences.

6.4.2 BLLNs, Maximum Nonparametric Likelihood, and Bayesian Maximum Probability

The Bayesian law of large numbers (BLLN) is a direct consequence of BST.

Bayesian Law of Large Numbers *Let $\Phi \subseteq \mathcal{P}(\mathcal{X})$ be a convex, closed set. Let $B(\hat{q}, \epsilon)$ be a closed ϵ-ball defined by the total variation metric and centered at the L-projection $\hat{q}$ of r on Φ. Then, for $\epsilon > 0$,*

$$\lim_{n\to\infty} \pi(q \in B(\hat{q}, \epsilon) \,|\, q \in \Phi, x_1^n; r) = 1, \quad \text{a.s. } r^\infty.$$

Thus, there is asymptotically *a posteriori* (a.s. r^∞) zero probability of a data-sampling distribution other than those arbitrarily close to the L-projection $\hat{q}$ of r on Φ.

BLLN is Bayesian counterpart of the CLLNs. When $\Phi = \mathcal{P}(\mathcal{X})$ the BLLN reduces to a special case, which is a counterpart of the law of large numbers. In this special case the L-projection $\hat{q}$ of the true data-sampling r on $\mathcal{P}(\mathcal{X})$ is just the data-sampling distribution r. Hence the BLLN can be in this case interpreted as indicating that, asymptotically, *a posteriori* the only possible data-sampling distributions are those that are arbitrary close to the "true" data-sampling distribution r.

The following example illustrates how BLLN, in the case where $\Phi \equiv \mathcal{P}(\mathcal{X})$, implies that the simplest problem of selecting of sampling distribution, has to be solved with the maximum nonparametric likelihood method. The SD problem is framed by the information-quadruple $(\mathcal{X}, \nu^n, \Phi, \pi(q))$. The objective is to select a sampling distribution from Φ.

Example 6.8
Let $\mathcal{X} = \{1, 2, 3, 4\}$, and let $r = [0.1, 0.4, 0.2, 0.3]$ be unknown to us. Let a random sample of size $n = 10^9$ be drawn from r, and let ν^n be the empirical measure that the sample induced. We assume that the mean of the true data-sampling distribution r is somewhere in the interval $[1, 4]$. Thus, r can be any pmf from $\mathcal{P}(\mathcal{X})$. Given the information $\mathcal{X}$, ν^n, $\Phi \equiv \mathcal{P}(\mathcal{X})$ and our prior $\pi(\cdot)$, the objective is to select a data-sampling distribution from Φ.

The problem presented in Example 6.8 is clearly an underdetermined, ill-posed inverse problem. Fortunately, BLLN regularizes it in the same way LLN did for the simplest empirical distribution selection problem, cf. Example 6.2 (Subsection 6.2.2). BLLN says that, given the sample, asymptotically *a posteriori* the only possible data-sampling distribution is the L-projection $\hat{q} \equiv r$ of r on $\Phi \equiv \mathcal{P}(\mathcal{X})$. Clearly, the true data-sampling distribution r is not known to us. Yet, for sufficiently large n, the sample-induced empirical measure ν^n is close to r. Hence, recalling BLLN, it is the L-projection of ν^n on Φ what we should select. Observe that this L-projection is just the probability distribution that maximizes $\sum_{i=1}^m \nu_i^n \log q_i$, the nonparametric likelihood.

We suggest the consistency requirement relative to potential methods for solving the SD problem. Namely, any method used to solve the problem should be such that it asymptotically conforms to the method implied by the Bayesian law of large numbers. We know that one such method is the maximum nonparametric likelihood. Another method that satisfies the consistency requirement and is more sound than MNPL, in the case of finite n, is

the method of maximum a posteriori probability (MAP), which selects

$$\hat{q}_{\text{MAP}} = \arg\sup_{q \in \Phi} \pi(q \mid \nu^n; r).$$

MAP, unlike MNPL, takes into account the prior distribution $\pi(q)$. It can be shown (cf. Grendár and Judge 2009a) that under the conditions for BLLN, MAP and MNPL asymptotically coincide and satisfy BLLN.

Although MNPL and MAP can legitimately be viewed as two different methods (and hence one should choose between them when n is finite), we prefer to view MNPL as an asymptotic instance of MAP (also known as Bayesian MaxProb), much like the view in (Grendár and Grendár 2001) that REM/MaxEnt is an asymptotic instance of the maximum probability method.

As CLLN regularizes ED problems, so does the Bayesian LLN for SD problems such as the one in Example 6.9.

Example 6.9
Let $\mathcal{X} = \{1, 2, 3, 4\}$, *and let* $r = [0.1, 0.4, 0.2, 0.3]$ *be unknown to us. Let a random sample of size* $n = 10^9$ *be drawn from* r, *and let* $\nu^n = [0.7, 0.42, 0.24, 0.27]$ *be the empirical measure that the sample induced. We assume that the mean of the true data-sampling distribution* r *is* 3.0; *i.e.,* $\Phi = \{q : \sum_{i=1}^{4} q_i x_i = 3.0\}$. *Note that the assumed value is different from the expected value of* X *under* r, 2.7. *Given the information* $\mathcal{X}$, ν^n, Φ *and our prior* $\pi(\cdot)$, *the objective is to select a data-sampling distribution from* Φ.

The BLLN prescribes the selection of a data-sampling distribution close to the L-projection $\hat{p}$ of the true data-sampling distribution r on Φ. Note that the L-projection of r on Φ, defined by linear moment consistency constraints $\Phi = \{q : \sum q(x_i)u_j(x_i) = a_j,\ j = 1, 2, \ldots, J\}$, where u_j is a real-valued function and $a_j \in \mathcal{R}$, belongs to the Λ-family of distributions (cf. Grendár and Judge 2009a),

$$\Lambda(r, u, \lambda, a) = \left\{ q : q(x) = r(x) \left[1 - \sum_{j=1}^{J} \lambda_j (u_j(x) - a_j) \right]^{-1}, x \in \mathcal{X} \right\}.$$

Since r is unknown to us, it is reasonable to replace r with the empirical measure ν^n induced by the sample X_1^n. Consequently, the BLLN instructs us to select the L-projection of ν^n on Φ, i.e., the data-sampling distribution that maximizes nonparametric likelihood. When n is finite, it is the maximum a posteriori probability data-sampling distribution(s) that should be selected. Thus, given certain technical conditions, BLLN provides a strong probabilistic justification for using the maximum a posteriori probability method and its asymptotic instance, the maximum nonparametric likelihood method, to solve the problem of selecting an SD.

Example 6.9 (cont'd)
Since n is sufficiently large, MNPL and MAP will produce a similar result. The L-projection $\hat{q}$ of ν^n on Φ belongs to the Λ family of distributions. The correct values $\hat{\lambda}$ of the parameters λ can be found by means of the convex dual problem (cf., e.g., Owen 2001):

$$\hat{\lambda} = \arg \inf_{\lambda \in \mathcal{R}^J} - \sum_i \nu_i^n \log \left(1 - \sum_j \lambda_j (u_j(x_i) - a_j) \right).$$

For the setting of Example 6.9, the L-projection $\hat{q}$ of ν^n on Φ can be found to be [0.043, 0.316, 0.240, 0.401].

6.4.3 Parametric SD Problem and Empirical Likelihood

Note that the SD problem is naturally in an empirical form. As such, there is only one step from the SD problem to the parametric SD problem, and this step means replacing Φ with a parametric set $\Phi(\Theta)$, where $\theta \in \Theta \subseteq \mathcal{R}^k$. The most common such set $\Phi(\theta)$, considered in *Econometrics*, is that defined by unbiased EEs, i.e., $\Phi(\Theta) = \bigcup_{\theta \in \Theta} \Phi(\theta)$, where

$$\Phi(\theta) = \left\{ q(x; \theta) : \sum_{i=1}^{m} q(x_i; \theta) u_j(x_i; \theta) = 0,\ j = 1, 2, \ldots, J \right\}.$$

The objective in solving the parametric SD problem is to select a representative sampling distribution(s) when only the information $(\mathcal{X}, \nu^n, \Phi(\Theta), \pi(q))$ is given. Provided that $\Phi(\Theta)$ is a convex, closed set and that n is sufficiently large, BLLN implies that the parametric Φ-problem should be solved with the maximum nonparametric likelihood method, i.e., by selecting

$$\hat{q}(\cdot; \theta) = \arg \inf_{q(\cdot;\theta) \in \Phi(\theta)} L(q(\cdot; \theta) \parallel \nu^n),$$

with $\theta = \hat{\theta}$, where

$$\hat{\theta}_{\mathrm{EL}} = \arg \inf_{\theta \in \Theta} L(\hat{q}(\cdot; \theta) \parallel \nu^n).$$

The resulting estimator $\hat{\theta}_{\mathrm{EL}}$ is known in the literature as the empirical likelihood (EL) estimator.

If n is finite/small, BLLN implies that the problem should be regularized with MAP method/estimator. It is worth highlighting that in the semiparametric EE setting, the prior $\pi(q)$ is put over $\Pi(\Theta)$, and the prior in turn induces a prior $\pi(\theta)$ over the parameter space Θ; cf. Florens and Rolin (1994).

BST and BLLN are also available for the case of continuous random variables; cf. (Grendár and Judge 2009a). In the case of EEs for continuous random variables, BLLN provides a consistency-under-misspecification argument for the continuous-form of EL estimator (see Equation (6.3)). BLLN also supports

the Bayesian MAP estimator

$$\hat{q}_{\text{MAP}}(x; \hat{\theta}_{\text{MAP}}) = \arg \sup_{q(x;\theta)\in\Phi(\theta)} \sup_{\theta\in\Theta} \pi(q(x;\theta) \mid x_1^n).$$

Since EL and the MAP estimators are consistent under misspecification, this provides a basis for the EL as well for the Bayesian MAP estimation methods. In conclusion it is worth stressing that in SD setting the other EMD estimators from the CR class (cf. Section 6.1) are not consistent, if the model is not correctly specified. The same holds, in general, for the posterior mean.

Example 6.10
As an illustration of application of EL in finance, consider a problem of estimation of the parameters of interest in rate diffusion models. In Lafférs (2009), parameters of Cox, Ingersoll, and Ross (1985) model, for an Euro overnight index average data, were estimated by empirical likelihood method, with the following set of estimating functions, for time t (Zhou 2001):

$$r_{t+1} - \mathrm{E}(r_{t+1} \mid r_t),$$
$$r_t[r_{t+1} - \mathrm{E}(r_{t+1} \mid r_t)],$$
$$\mathrm{V}(r_{t+1} \mid r_t) - [r_{t+1} - \mathrm{E}(r_{t+1} \mid r_t)]^2,$$
$$r_t\{\mathrm{V}(r_{t+1} \mid r_t) - [r_{t+1} - \mathrm{E}(r_{t+1} \mid r_t)]^2\}.$$

There, r_t denotes the interest rate at time t, V denotes the variance. In Lafférs (2009) also a Monte Carlo study of small sample properties of EL estimator was conducted; cf. also Zhou (2001).

6.5 Summary

The Empirical Minimum Divergence (EMD) approach to estimation and inference, described in Section 6.1, is an attractive alternative to the generalized method of Moments. EMD comprises two components: a parametric model, which is usually specified by means of EEs, and a divergence (discrepancy) measure of a pdf with respect to the true sampling distribution. The divergence is minimized among parametrized pdf's from the model set, and this way a pdf is selected. The selected parametrized pdf depends on the true, yet unknown in practice, sampling distribution. Since the assumed discrepancy measures are convex and the model set is a convex set, the optimization problem has its convex dual equivalent formulation; cf. Equation 6.1. The convex dual problem (Equation 6.1) can be tied to the data by replacing the expectation by its empirical analogue; cf. (Equation 6.2). This way the data are taken into account and the EMD estimator results.

A researcher can choose between two possible ways of using the parametric model, defined by EEs. One option is to use the EEs to define a feasible set $\Phi(\Theta)$ of possible parametrized sampling distributions. Then the objective of EMD procedure is to select a parametrized sampling distribution (SD) from the model set $\Phi(\Theta)$, given the data. This modeling strategy and the objective deserve a name, and we call it the parametric SD problem. The other option is to let the EEs define a feasible set $\Pi(\Theta)$ of possible parametrized empirical distributions and use the observed, data-based empirical pmf in place of a sampling distribution. If this option is followed, then, given the data, the objective of the EMD procedure is to select a parametrized empirical distribution from the model set $\Pi(\Theta)$, given the data; we call it the parametric empirical ED problem. The empirical attribute stems for the fact that the data are used to estimate the sampling distribution.

In addition to the possibility of choosing between the two strategies, a researcher who follows the EMD approach to estimation and inference can select a particular divergence measure. Usually, divergence measures from Cressie–Read (CR) family are used in the literature. Prominent members of the CR-based class of EMD estimators are: maximum empirical likelihood estimator (MELE), empirical maximum maximum entropy estimator (EMME), and Euclidean empirical likelihood (EEL) estimator. Properties of EMD estimators have been studied in numerous works. Of course, one is not limited to the "named" members of CR family. Indeed, in the literature an option of letting the data select "the best" member of the family, with respect to a particular loss function, has been explored.

Consistency is perhaps the least debated property of estimation methods. EMD estimators are consistent, provided that the model is well-specified; i.e., the feasible set (being it Φ or Π) contains the true data-sampling distribution r. However, models are rarely well-specified. It is thus of interest to know which of the EMD methods of information recovery is consistent under misspecification. And here the large deviations (LD) theory enters the scene. LD theory helps to both define consistency under misspecification and to identify methods with this property. Large deviations are rather a technical subfield of the probability theory. Our objective has been to provide a nontechnical introduction to the basic theorems of LD, and step-by-step show the meaning of the theorems for consistency-under-misspecification requirement.

Since there are two modeling strategies, there are also two sets of LD theorems. LD theorems for empirical measures are at the base of classic (orthodox) LD theory. The theorems suggest that the relative entropy maximization method (REM, aka MaxEnt) possesses consistency-under-misspecification in the nonparametric form of the ED problem. The consistency extends also to the empirical parametric ED problem, where it is the empirical maximum maximum entropy method that has the desired property. LD theorems for sampling distributions are rather recent. They provide a consistency-under-misspecification argument in favor of the Bayesian maximum a posteriori probability, maximum nonparametric likelihood, and empirical likelihood

methods in nonparametric and semiparametric form of the SD problem, respectively.

6.6 Notes on Literature

1. The LD theorems for empirical measures discussed here can be found in any standard book on LD theory. We recommend Dembo and Zeitouni (1998), Ellis 2005, Csiszár (1998), and Csiszár and Shields (2004) for readers interested in LD theory and closely related method of types, which is more elucidating. An accessible presentation of ST and CLLN can be found in Cover and Thomas (1991). Proofs of the theorems cited here can be found in any of these sources. A physics-oriented introduction to LD can be found in Aman and Atmanspacher (1999) and Ellis (1999).
2. Sanov theorem (ST) was considered for the first time in Sanov (1957), extended by Bahadur and Zabell (1979). Groeneboom, Oosterhoff, and Ruymgaart (1979) and Csiszár (1984) proved ST for continuous random variables; cf. Csiszár (2006) for a lucid proof of continuous ST. Csiszár, Cover, and Choi (1987) proved ST for Markov chains. Grendár and Niven (2006) established ST for the Pólya urn sampling. The first form of CLLNs known to us is that of Bártfai (1972). For developments of CLLN see Vincze (1972), Vasicek (1980), van Campenhout and Cover (1981), Csiszár (1984,1985,1986), Brown and Smith (1986), Harremoës (2007), among others.
3. Gibbs conditioning principle (GCP) (cf. Csiszár 1984; Lanford 1973), and (see also Csiszár 1998; Dembo and Zeitouni 1998), which was not discussed in this chapter, is a stronger LD result than CLLN. GCP reads:

   Gibbs conditioning principle: *Let $\mathcal{X}$ be a finite set. Let Π be a closed, convex set. Let $n \to \infty$. Then, for a fixed t,*

   $$\lim_{n\to\infty} \pi(X_1 = x_1, \ldots, X_t = x_t \mid \nu^n \in \Pi; q) = \prod_{l=1}^{t} \hat{p}_{x_l}.$$

   Informally, GCP says that, if the sampling distribution q is confined to produce sequences which lead to types in a set Π, then elements of any such sequence of fixed length t will behave asymptotically conditionally as if they were drawn identically and independently from the I-projection $\hat{p}$ of q on Π — provided that the last is unique. There is no direct counterpart of GCP in the Bayesian Φ-problem setting. In order to keep symmetry of the exposition, we decided to not discuss GCP in detail.
4. Jaynes' views of maximum entropy method can be found in Jaynes (1989). In particular, the entropy concentration theorem (cf. Jaynes 1989)

is worth mentioning. It says, using our notation, that, as $n \to \infty$, $2n\triangle H(\nu^n) \sim \chi^2_{m-J-1}$ and $H(p) = -\sum p_i \log p_i$ is the Shannon entropy.

For a mathematical treatment of the maximum entropy method see Csiszár (1996, 1998). Various uses of MaxEnt are discussed in Solana-Ortega and Solana (2005). For a generalization of MaxEnt which is of direct relevance to *Econometrics*, see Golan, Judge, and Miller (1996), and also Golan (2008).

Maximization of the Tsallis entropy (MaxTent) leads to the same solution as maximization of Rényi entropy. Bercher proposed a few arguments in support of MaxTent; cf. Bercher (2008) for a survey.

For developments of the maximum probability method cf. Boltzmann (1877), Vincze (1972), Vincze (1997), Grendár and Grendár (2001), Grendár and Grendár (2004), Grendár and Niven (2006), Niven (2007). For the asymptotic connection between MaxProb and MaxEnt see Grendár and Grendár (2001, 2004).

5. While the LD theorems for empirical measures have already found their way into textbooks, discussions of LD for data-sampling distributions are rather recent. To the best of our knowledge, the first Bayesian posterior convergence via LD was established by Ben-Tal, Brown, and Smith (1987). In fact, their Theorem 1 covers a more general case where it is assumed that there is a set of empirical measures rather than a single such a measure ν^n. The authors extended and discussed their results in Ben-Tal, Brown, and Smith (1988). For some reasons, these works remained overlooked. More recently, ST for data-sampling distributions was established in an interesting work by Ganesh and O'Connell (1999). The authors established BST for finite $\mathcal{X}$ and well-specified model. In Grendár and Judge (2009a), Bayesian ST and the Bayesian LLN were developed for $\mathcal{X} = \mathcal{R}$ and a possibly misspecified model.
6. Relevance of LD for empirical measures for empirical estimator choice was recognized by Kitamura and Stutzer (1997), where LD justification of empirical MaxMaxEnt was discussed.
7. Finding empirical likelihood or empirical MaxMaxEnt estimators is a demanding numeric problem; cf., e.g., Mittelhammer and Judge (2001). In Brown and Chen (1998) an approximation to EL via the Euclidean likelihood was suggested, which makes the computations easier. Chen, Variyath, and Abraham (2008) proposed the Adjusted EL which mitigates a part of the numerical problem of EL. Recently, it was recognized that empirical likelihood and related methods are susceptible to the empty set problem that requires a revision of the available empirical evidence on EL-like methods; cf. Grendár and Judge (2009b).
8. Properties of estimators from EMD class were studied in numerous works; cf. Back and Brown (1990), Baggerly (1998), Baggerly (1999), Bickel et al. (1993), Chen et al. (2008), Corcoran (2000), DiCiccio, Hall, and Romano (1991), DiCiccio, Hall, and Romano (1990), Grendár and Judge (2009a), Imbens (1993), Imbens, Spady, and Johnson (1998), Jing and Wood (1996), Judge and Mittelhammer (2004), Judge and

Mittelhammer (2007), Kitamura and Stutzer (1997), Kitamura and Stutzer (2002), Lazar (2003), Mittelhammer and Judge (2001), Mittelhammer and Judge (2005), Mittelhammer, Judge, and Schoenberg (2005), Newey and Smith (2004), Owen (1991), Qin and Lawless (1994), Schennach (2005), Schennach (2004), Schennach (2007), Grendár and Judge (2009a), Grendár and Judge (2009b), among others.

6.7 Acknowledgments

Valuable feedback from Doug Miller, Assad Zaman, and an anonymous reviewer is gratefully acknowledged.

References

Amann, A., and H. Atmanspacher. 1999. Introductory remarks on large deviations statistics. *J. Sci. Explor.* 13(4):639–664.

Back, K., and D. Brown. 1990. Estimating distributions from moment restrictions. Working paper, Graduate School of Business, Indiana University.

Baggerly, K. A. 1998. Empirical likelihood as a goodness-of-fit measure. *Biometrika*. 85(3):535–547.

Baggerly, K. A. 1999. Studentized empirical likelihood and maximum entropy. Technical report, Rice University, Dept. of Statistics, Houston, TX.

Bahadur, R., and S. Zabell. 1979. Large deviations of the sample mean in general vector spaces. *Ann. Probab*. 7:587–621.

Bártfai, P. 1972. On a conditional limit theorem. *Coll. Math. Soc. J. Bolyai*. 9:85–91.

Ben-Tal, A., D. E., Brown, and R. L. Smith. 1987. Posterior convergence under incomplete information. Technical report 87–23. University of Michigan, Ann Arbor.

Ben-Tal, A., D. E., Brown, and R. L. Smith. 1988. Relative entropy and the convergence of the posterior and empirical distributions under incomplete and conflicting information. Technical report 88–12. University of Michigan Ann Arbor.

Bercher, J.-F. 2008. Some possible rationales for Rényi-Tsallis entropy maximization. In *International Workshop on Applied Probability, IWAP 2008*.

Bickel, P. J., C. A. J., Klassen, Y., Ritov, and J. Wellner. 1993. *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Boltzmann, L. 1877. Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheilichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht. *Wiener Berichte* 2(76):373–435.

Brown, B. M., and S. X. Chen. 1998. Combined and least squares empirical likelihood. *Ann. Inst. Statist. Math.* 90:443–450.

Brown, D. E., and R. L. Smith. 1986. A weak law of large numbers for rare events. Technical report 86–4. University of Michigan, Ann Arbor.

Chen, J., A. M., Variyath, and B. Abraham. 2008. Adjusted empirical likelihood and its properties. *J. Comput. Graph. Stat*. 17(2):426–443.

Corcoran, S. A. 2000. Empirical exponential family likelihood using several moment conditions. *Stat. Sinica*. 10:545–557.

Cover, T., and J. Thomas. 1991. *Elements of Information Theory*. New York: Wiley.
Cox, J. C., J. E., Ingersoll, and S. A. Ross. 1985. A theory of the term structure of interest rates. *Econometrica* 53:385–408.
Cox, S. J., G. J., Daniell, and D. A. Nicole. 1998. Using maximum entropy to double ones expected winnings in the UK National Lottery. *JRSS Ser. D*. 47(4):629–641.
Cressie, N., and T. Read. 1984. Multinomial goodness of fit tests. *JRSS Ser. B*. 46:440–464.
Cressie, N., and T. Read. 1988. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.
Csiszár, I. 1984. Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.* 12:768–793.
Csiszár, I. 1985. An extended maximum entropy principle and a Bayesian justification theorem. In *Bayesian Statistics 2*, 83–98. Amsterdam: North-Holland.
Csiszár I. 1996. MaxEnt, mathematics and information theory. In *Maximum Entropy and Bayesian Methods*. K. M. Hanson and R. N. Silver (eds.), pp. 35–50. Dordrecht: Kluwer Academic Publishers.
Csiszár I. 1998. The method of types. *IEEE IT*. 44(6):2505–2523.
Csiszár, I. 2006. A simple proof of Sanov's theorem. *Bull. Braz. Math. Soc*. 37(4):453–459.
Csiszár, I., T., Cover, and B. S. Choi. 1987. Conditional limit theorems under Markov conditioning, *IEEE IT*. 33:788–801.
Csiszár, I., and P. Shields. 2004. Information theory and statistics: a tutorial. *Found. Trends Comm. Inform. Theory*. 1(4):1–111.
Dembo, A., and O. Zeitouni. 1998. *Large Deviations Techniques and Applications*. New York: Springer-Verlag.
DiCiccio, T. J., P. J. Hall, and J. Romano. 1990. Nonparametric confidence limits by resampling methods and least favorable families. *I.S.I. Review*. 58:59–76.
DiCiccio, T. J., P. J. Hall, and J. Romano. 1991. Empirical likelihood is Bartlett-correctable. *Ann. Stat.* 19:1053–1061.
Ellis, R. S. 1999. The theory of large deviations: from Boltzmann's 1877 calculation to equilibrium macrostates in 2D turbulence. *Physica D*. 106–136.
Ellis, R. S. 2005. *Entropy, Large Deviations and Statistical Mechanics*. 2nd ed. New York: Springer-Verlag.
Farrell, L., R., Hartley, G., Lanot, and I. Walker. 2000. The demand for Lotto: the role of conscious selection. *J. Bus. Econ. Stat*. 18(2):228–241.
Florens, J.-P., and J.-M. Rolin. 1994. Bayes, bootstrap, moments. Discussion paper 94.13. Institute de Statistique, Université catholique de Louvain.
Ganesh, A., and N. O'Connell. 1999. An inverse of Sanov's Theorem. *Stat. Prob. Lett*. 42:201–206.
Ghosal, A., J. K., Ghosh, and R. V. Ramamoorthi. 1999. Consistency issues in bayesian nonparametrics. In *Asymptotics, Nonparametrics and Time Series: A Tribute to Madan Lal Puri*, pp. 639–667. New York: Marcel Dekker.
Ghosh, J. K., and R. V. Ramamoorthi. 2003. *Bayesian Nonparametrics*. New York: Springer-Verlag.
Godambe, V. P., and B. K. Kale. 1991. Estimating functions: an overview. In *Estimating Functions*. V. P. Godambe (ed.), pp. 3–20. Oxford, U.K.: Oxford University Press.
Golan, A. 2008. Information and entropy econometrics: a review and synthesis. *Foundations and Trends in Econometrics*, 2(12):1–145.
Golan, A., G., Judge, and D. Miller. 1996. *Maximum Entropy Econometrics. Robust Estimation with Limited Data*. New York: Wiley.
Grendár M. Jr., and M. Grendár. 2001. What is the question that MaxEnt answers? A probabilistic interpretation. In *Bayesian Inference and Maximum Entropy Methods in*

Science and Engineering. A. Mohammad-Djafari (ed.), pp. 83-94. Melville, NY: AIP. Online at arxiv:math-ph/0009020.

Grendár. M., Jr., and M. Grendár. 2004. Asymptotic identity of μ-projections and I-projections. *Acta Univ. Belii. Math.* 11:3–6.

Grendár, M., and G. Judge. 2008. Large deviations theory and empirical estimator choice. *Econometric Rev.* 27(4–6):513–525.

Grendár, M., and G. Judge. 2009a. Asymptotic equivalence of empirical likelihood and Bayesian MAP. *Ann. Stat.* 37(5A):2445–2457.

Grendár, M., and G. Judge. 2009b. Empty set problem of maximum empirical likelihood methods. *Electron. J. Stat.* 3:1542–1555.

Grendár, M., and R. K. Niven. 2006. The Pólya urn: limit theorems, Pólya divergence, maximum entropy and maximum probability. On-line at: arXiv:cond-mat/0612697.

Groeneboom, P., J., Oosterhoff, and F. H. Ruymgaart. 1979. Large deviation theorems for empirical probability measures. *Ann. Probab.* 7:553–586.

Hall, A. R. 2005. *Generalized Method of Moments*. Advanced Texts in Econometrics. Oxford, U.K.: Oxford University Press.

Hansen, L. P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50:1029–1054.

Harremoës, P. 2007. Information topologies with applications. In *Entropy, Search and Complexity*, I. Csiszár et al. (eds.), pp.113–150. New York: Springer.

Imbens, G. W. 1993. A new approach to generalized method of moments estimation. Harvard Institute of Economic Research working paper 1633.

Imbens, G. W., R. H., Spady, and P. Johnson. 1998. Information theoretic approaches to inference in moment condition models. *Econometrica* 66(2):333–357.

Jaynes, E. T. 1989. *Papers on Probability, Statistics and Statistical Physics*. 2nd ed. R. D. Rosenkrantz (ed.). New York: Springer.

Jing, B.-Y., and T. A. Wood. 1996. Exponential empirical likelihood is not Bartlett correctable. *Ann. Stat.* 24:365–369.

Jones, L. K., and C. L. Byrne. 1990. General entropy criteria for inverse problems, with applications to data compression, pattern classification and cluster analysis. *IEEE IT* 36(1):23–30.

Judge G. G., and R. C. Mittelhammer. 2004. A semiparametric basis for combining estimation problems under quadratic loss. *JASA* 99:479–487.

Judge, G. G., and R. C. Mittelhammer. 2007. Estimation and inference in the case of competing sets of estimating equations. *J. Econometrics* 138:513–531.

Kitamura, Y. 2006. Empirical likelihood methods in econometrics: theory and practice. In *Advances in Economics and Econometrics: Theory and Applications, Ninth world congress*. Cambridge, U.K.: CUP.

Kitamura, Y., and M. Stutzer. 1997. An information-theoretic alternative to generalized method of moments estimation. *Econometrica* 65:861–874.

Kitamura, Y., and M. Stutzer. 2002. Connections between entropic and linear projections in asset pricing estimation. *J. Econometrics* 107:159–174.

Laffér s, L. 2009. Empirical likelihood estimation of interest rate diffusion model. Master's thesis, Comenius University.

Lanford, O. E. 1973. Entropy and equilibrium states in classical statistical mechanics. In *Statistical Mechanics and Mathematical Problems*, A. Lenard (ed.), LNP 20, pp. 1–113. New York: Springer.

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

Mittelhammer, R. C., and G. G. Judge. 2001. Robust empirical likelihood estimation of models with non-orthogonal noise components. *J. Agricult. Appl. Econ.* 35: 95–101.

Mittelhammer, R. C., and G. G. Judge. 2005. Combining estimators to improve structural model estimation and inference under quadratic loss. *J. Econometrics* 128(1):1–29.

Mittelhammer, R. C., Judge, G. G., and D. J. Miller. 2000. *Econometric Foundations*. Cambridge, U.K.: CUP.

Mittelhammer, R. C., Judge, G. G., and R. Schoenberg. 2005. Empirical evidence concerning the finite sample performance of EL-type structural equations estimation and inference methods. In *Identification and Inference for Econometric Models. Essays in Honor of Thomas Rothenberg*. D. Andrews, and J. Stock (eds.). Cambridge, U.K.: Cambridge University Press.

Newey, W., and R. J. Smith. 2004. Higher-order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 72:219–255.

Niven, R. K. 2007. Origins of the combinatorial basis of entropy. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. K. H. Knuth et al. (eds.). pp. 133–142. Melville, NY: AIP.

Owen, A. B. 1991. Empirical likelihood for linear models. *Ann. Stat.* 19:1725–1747.

Owen, A. B. 2001. *Empirical Likelihood*. New York: Chapman-Hall/CRC.

Qin, J., and J. Lawless. 1994. Empirical likelihood and general estimating equations. *Ann. Stat.* 22:300–325.

Sanov, I. N. 1957. On the probability of large deviations of random variables. *Mat. Sbornik*. 42:11–44. (in Russian).

Schennach, S. M. 2004. Exponentially tilted empirical likelihood. Working paper, Department of Economics, University of Chicago.

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

Schennach, S. M. 2007. Point estimation with exponentially tilted empirical likelihood. *Ann. Stat.* 35(2):634–672.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell Sys. Tech. J.* 27:379–423 and 27:623–656.

Solana-Ortega, A., and V. Solana. 2005. Entropic inference for assigning probabilities: some difficulties in axiomatics and applications, In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. A. Mohammad-Djafari (ed.). pp. 449–458. Melville, NY: AIP.

van Campenhout J. M., and T. M. Cover. 1981. Maximum entropy and conditional probability. *IEEE IT* 27:483–489.

Vasicek O. A. 1980. A conditional law of large numbers. *Ann. Probab.* 8:142–147.

Vincze, I. 1972. On the maximum probability principle in statistical physics. *Coll. Math. Soc. J. Bolyai*. 9:869–893.

Vincze, I. 1997. Indistinguishability of particles or independence of the random variables? *J. Math. Sci.* 84:1190–1196.

Walker. S. 2004. New approaches to bayesian consistency. *Ann. Stat.* 32:2028–2043.

Walker, S., A., Lijoi, and I. Prünster. 2004. Contibutions to the understanding of bayesian consistency. Working paper no. 13/2004, International Centre for Economic Research, Turin.

Zhou, H. 2001. Finite sample properties of EMM, GMM, QMLE, and MLE for a square-root interest rate diffusion model. *J. Comput. Finance* 5:89–122.