NONPARAMETRIC LIKELIHOOD: EFFICIENCY AND ROBUSTNESS

YUICHI KITAMURA*

Abstract. Nonparametric likelihood is a natural generalization of the parametric maximum likelihood estimation (MLE) procedure, which has been the workhorse in empirical economics. An interesting fact is that the MLE procedure remains valid in many stochastic models that have nonparametric components, provided that the "likelihood function" is formulated appropriately. Such a likelihood function can be obtained by using multinomial distribution functions, supported by observed data values, to approximate the underlying distribution nonparametrically. This yields the so-called nonparametric likelihood function, and its maximizer is often labeled as the nonparametric maximum likelihood estimator (NPMLE). This is an extension of great interest, since econometric theory rarely suggests a parametric form of the probability law of data. Being a nonparametric method, nonparametric likelihood is robust against misspecification. At the same time, it often achieves good properties that are analogous to those of parametric likelihood. The purpose of this paper is to provide an overview of the growing literature of nonparametric likelihood, and to discuss new theoretical developments in this area of research. Various applications of nonparametric likelihood are explored, with some emphasis on the analysis of biased samples and data combination problems.

1. Introduction

Nonparametric likelihood is a natural generalization of the parametric maximum likelihood estimation (MLE) procedure, which has been the workhorse in empirical economics. Desirable properties of the likelihood method in a parametric model, where the probability law of observations is determined by a model that is known to the researcher up to a finite dimensional parameter, are well-known: it provides a consistent estimator and enjoys various optimality properties under general regularity conditions. Also, the MLE procedure essentially estimate the entire probability law of the stochastic model, which is useful for policy analysis.

An interesting fact is that the MLE procedure remains valid in many stochastic models that have nonparametric components, provided that the "likelihood function" is formulated appropriately. It can be carried out by using approximating distribution functions, supported by observed data values,

Date: November 12, 2006.

*This paper is prepared for the 2006 Nakahara Prize Lecture at the 2006 Fall Meeting of the Japanese Economic Society. I acknowledge financial support from the National Science Foundation via grants SES-0241770 and SES-0551271.

to estimate the underlying distribution function nonparametrically. This yields the nonparametric likelihood function, and its maximizer is often labeled as the nonparametric maximum likelihood estimator (NPMLE). This is an extension of great interest, since econometric theory rarely suggests a parametric form of the probability law of data. Being a nonparametric method, nonparametric likelihood is robust against misspecification unlike the conventional fully parametric procedure. At the same time, it often achieves good properties akin to those of parametric likelihood. The literature on nonparametric likelihood has been growing rapidly in the recent literature, partly due to the discovery of empirical likelihood (EL) by Owen (1988).

The purpose of this paper is to provide an overview of the growing literature of nonparametric likelihood, and to discuss new theoretical developments in this area of research. Various applications of the NPMLE are explored, with some emphasis on the analysis of biased samples and data combination problems.

The following is a (selective) list of useful features of nonparametric likelihood:

(1) The use of empirically observed values as the support of approximating distributions has an intuitive appeal, and requires minimal *a priori* information about the underlying distribution. In addition to being fully nonparametric, the NPMLE procedure demands little restriction about the smoothness and other features of the unknown true distribution. This is an attractive feature that makes nonparametric likelihood flexible and robust. A related fact is that it often sidesteps the choice of tuning parameters such as bandwidth in kernel smoothing. This is important in practice, since many nonparametric procedure suffer from arbitrariness due to their dependence of the choice of tuning parameters.

(2) Another aspect of the NPMLE procedure that is equally important in its empirical applications is its efficiency. This paper takes a rather broad view of efficiency concepts, including the classical notion of semiparametric efficiency (Bickel, Klassen, Ritov, and Wellner (1993)), large deviations efficiency of estimators and tests, and higher order efficiency. Methods based on NPMLE achieve efficiency properties in many applications. These may be regarded as extensions of efficiency properties of the parametric likelihood procedure, though in some cases subtle arguments are required to prove the efficiency of NPMLE.

(3) There is an additional feature that NPMLE shares with parametric MLE. Many semiparametric procedures tend to focus on estimation of the parameters of interest. When it comes to policy analysis, however, it is often useful to have the entire probability law of the model estimated. The NPMLE offers a natural estimator of the full stochastic model. An example is a moment condition model, to which both GMM and EL apply. GMM focuses on estimating the (finite dimensional) parameter in the moment function. EL, on the other hand, estimates the underlying distribution and the parameters in the moment function jointly, thereby yielding efficient estimators for both of them.

(4) Combining different data sets to achieve identification and/or efficiency is an important topic in the recent econometric literature. See Ridder and Moffitt (2003) for a comprehensive and updated survey on this topic. NPMLE offers a natural way to carry out data combination; see Section 2.5.

The subsequent sections explore the above topics. There are some materials in the paper that have been discussed by the survey chapters such as Cosslett (1993), Cosslett (1997)), Owen (2001), Kitamura (2006a) and Kitamura (2006b). The aim of the paper is to illustrate the fundamental concept of nonparametric likelihood, demonstrate its efficacy in various contexts, present some algorithms, and discuss issues that deserve further research. It does not, therefore, offer mathematically rigorous treatments, for which the readers are referred to the original papers.

2. Applications of NPMLE

The nonparametric likelihood method applies to a wide variety of models that are important in empirical economics. Before introducing specific models, however, it is beneficial to observe that the empirical distribution of a random sample is the NPMLE in the absence of a model. Suppose the econometrician observes $\{z_i\}_{i=1}^n$, a random sample from a CDF F_0, which is unknown. Let Δ denote the simplex $\{(p_1, ..., p_n) : \sum_{i=1}^n p_i = 1, 0 \leq p_i, i = 1, ..., n\}$. Use the vector $(p_1, ..., p_n) \in \Delta$ to "parameterize" the unknown distribution F_0 by $F_n(z) = \sum_{i=1}^n p_i 1\{z_i \leq z\}$, $z \in \mathbb{R}$, where $1\{\cdot\}$ signifies the usual indicator function. The discrete distribution F_n is used to approximate the true F_0. It has density function $f_n(z) = \sum_{i=1}^n p_i 1\{z = z_i\}$ with respect to the counting measure on the

observed values $\{z_i\}_{i=1}^n$, yielding the nonparametric log-likelihood function

$$\ell = \sum_{i=1}^n \log p_i, \quad (p_1, ..., p_n) \in \Delta.$$

Maximizing this with respect to $(p_1, ..., p_n)$ yields $\hat{p}_i = \frac{1}{n}, i = 1, ..., n$. Evaluating F_n at these values, obtain

$$\widehat{F}_n(z) = \frac{1}{n}\sum_{i=1}^n 1\{z_i \leq z\} = \frac{1}{n}\#\{z_i : z_i \leq z\},$$

which is the empirical distribution function. It is known that the empirical distribution function has certain desirable asymptotic properties (see, for example, Dvoretzky, Kiefer, and Wolfowitz (1956)).

2.1. **Moment Condition Model. Empirical Likelihood.** Let z denote a random variable and F_0 its CDF. Suppose the expectation of an $\mathbb{R}^q$-valued function $g(z, \theta_0)$, which is known up to the finite dimensional parameter θ_0 in $\Theta \subset \mathbb{R}^k$, is restricted to be zero:

$$E[g(z, \theta_0)] = \int g(z, \theta_0) dF_0(z) = 0. \tag{2.1}$$

The unknown distribution F_0 is the nonparametric component of this model. The econometrician observes iid data $\{z_i\}_{i=1}^n$, where z_i obeys the above model. Owen (1988) and subsequent papers considered applications of the NPMLE to the moment condition model. Using the notation used above in describing the empirical distribution, let $F_n(z) = \sum_{i=1}^n p_i 1\{z_i \leq z\}$, $z \in \mathbb{R}$ denote an approximating distribution for F_0 parameterized by $\{(p_1, ..., p_n) : \sum_{i=1}^n p_i = 1, 0 \leq p_i, i = 1, ..., n\}$. The nonparametric log-likelihood function to be maximized is

$$\ell(p_1, ..., p_n, \theta) = \sum_{i=1}^n \log p_i, \quad \sum_{i=1}^n g(z_i, \theta) p_i = 0, \quad (p_1, ..., p_n) \in \Delta, \theta \in \Theta.$$

Let $(\hat{\theta}_{\mathrm{EL}}, \hat{p}_{\mathrm{EL}1}, ..., \hat{p}_{\mathrm{EL}n})$ denote the value of $(\theta, p_1, ..., p_n) \in \Theta \times \Delta$ that maximizes ℓ. This is called the (maximum) empirical likelihood estimator. The NPMLE for θ_0 and F_0 are $\hat{\theta}_{\mathrm{EL}}$ and $\widehat{F}_{\mathrm{EL}} = \sum_{i=1}^n \hat{p}_{\mathrm{EL}i} 1\{z_i \leq z\}$.

Remark 2.1. One might expect that the high dimensionality of the parameter space $\Theta \times \Delta$ makes the above maximization problem intractable for any practical application. Fortunately, that is not the case, if one uses the following nested procedure. First, fix θ at a value in Θ and consider the log-likelihood with the parameters $(p_1, ..., p_n)$ "profiled out":

$$\ell(\theta) = \sup_{(p_1, ..., p_n) \in \Delta} \ell(p_1, ..., p_n, \theta) \text{ subject to } \sum_{i=1}^n p_i g(z_i, \theta) = 0. \tag{2.2}$$

A straightforward application of the Lagrange multiplier method shows that $\ell(\theta)$ is represented by

$$\ell(\theta) = \min_{\gamma \in \mathbb{R}^q} - \sum_{i=1}^{n} \log(1 + \gamma' g(z_i, \theta))) - n \log n \tag{2.3}$$

(see, for example, Kitamura (2006b)). The numerical evaluation of the function $\ell(\cdot)$ is easy, because (2.3) is a low dimensional convex maximization problem, for which a simple Newton algorithm works. Second, obtain the empirical likelihood estimator $\hat{\theta}_{\mathrm{EL}}$ as the maximizer of (2.3). The maximization of $\ell(\theta)$ with respect to θ is typically carried our using a nonlinear optimization algorithm.

Remark 2.2. Basic properties of the empirical likelihood procedure are now well-understood. The EL estimator $\hat{\theta}_{\mathrm{EL}}$ is $n^{1/2}$-consistent and asymptotically normal. Let D and S denote $E[\nabla_\theta g(z, \theta_0)]$ and $E[g(z, \theta_0) g(z, \theta_0)']$, then its asymptotic distribution is given by $\mathrm{N}(0, (D'SD)^{-1})$. Also, suppose R is a known $\mathbb{R}^s-$valued function of θ, and the econometrician poses a hypothesis that θ_0 is restricted as $R(\theta_0) = 0$, where the s restrictions are independent. This can be tested by forming a nonparametric analog of the parametric likelihood ratio statistic. Let $r = -2\left(\sup_{\theta: R(\theta)=0} \ell(\theta) - \sup_{\theta \in \Theta} \ell\right)$, then this obeys the chi-square distribution with s degrees of freedom asymptotically under the null. The factor r is called the empirical likelihood ratio (ELR) statistic. ELR also applies to testing overidentifying restrictions: see Section 4. These properties and other basics of EL and related methods have been studied extensively in the literature: see Qin and Lawless (1994), Imbens (1997), Kitamura (1997), Kitamura and Stutzer (1997), Qin and Lawless (1994), Smith (1997), Imbens, Spady, and Johnson (1998), and Newey and Smith (2004).

2.2. **Conditional Moment Restriction Model.** Suppose economic theory implies that the conditional mean of $g(z, \theta_0)$ given a vector of covariates x is zero:

$$E[g(z, \theta_0)|x] = 0. \tag{2.4}$$

This restriction is stronger than (2.1). Though one can choose an arbitrary function of x as an instrument, this can be problematic since (1) choosing an instrument that delivers strong identification may be a difficult task, and (2) an arbitrary instrument does not achieve efficiency in general. Kitamura, Tripathi, and Ahn (2004) propose a technique to incorporate the information in the conditional moment restriction into empirical likelihood.

The econometrician observes random sample $\{(x_i, z_i)\}_{i=1}^n$ where each (x_i, z_i) satisfies (2.4). Let $F_0(z|x)$ and $H_0(x)$ denote the conditional distribution of z_i given $x_i = x$, and the marginal distribution of x_i, respectively. The model (2.4) imposes restrictions on $F_0(z|x)$. It is therefore appropriate

to apply the NPMLE procedure to treat $F_0(z|x)$ nonparametrically. Let the discrete distributions $F_n(z|x_i) = \sum_{j=1}^n p_{ij} 1\{z_j \leq z\}, i = 1, ..., n$ and $H_n(x) = \sum_{i=1}^n p_i 1\{x_i \leq x\}$ be approximations for $F_0(z|x)$ and $H_0(x)$. The parameters $\{\{p_{ij}\}_{j=1}^n\}_{i=1}^n$, $\{p_i\}_{i=1}^n$ and θ are to be determined by the data. These parameters need to satisfy, under (2.4), the following conditions

$$\sum_{j=1}^n p_{ij} g(z_j, \theta) = 0, i = 1, ..., n, \theta \in \Theta, \tag{2.5}$$

in addition to the obvious restrictions

$$\sum_{i=1}^n p_{ij} = 1, \quad p_{ij} \geq 0, \quad \sum_{i=1}^n p_i = 1, \quad p_i \geq 0 \quad \text{for } 1 \leq i, j \leq n. \tag{2.6}$$

If the covariate x is known to have finite support, the NPMLE is straightforward to define. Let the support points $\mathbf{x}(1), ..., \mathbf{x}(K)$, and $\{\{p_{kj}\}_{j=1}^n\}_{k=1}^K$ and $\{p_k\}_{k=1}^K$ denote the conditional probabilities and the marginal probabilities as above. The nonparametric log-likelihood function is:

$$\ell = \sum_{i=1}^n \ell_i, \quad \ell_i = \sum_{k=1}^K 1\{x_i = \mathbf{x}(k)\}(\log p_{ki} + \log p_i), \tag{2.7}$$

i.e., the summation of log-likelihood contributions $l_i, i = 1, ..., n$. The more interesting case is the model (2.4) with continuously distributed x, however. If this is the case, the above approach fails: For one thing, $p_{ij}, j \neq i$ would not show up in the likelihood function. Kitamura, Tripathi, and Ahn (2004) utilize the kernel regression technique to resolve this problem. Let K and h be an appropriate kernel function and a bandwidth parameter that goes to zero as n approaches infinity. The idea is to replace the indicator function in (2.7) with a Nadaraya-Watson kernel weighted average:

$$\sum_{j=1}^n w_{ij} \log p_{ij}, \quad w_{ij} = \frac{K(\frac{x_i - x_j}{h})}{\sum_{j=1}^n K(\frac{x_i - x_j}{h})}$$

to use observations with the values of x close to x_i in calculating the i-th log-likelihood contribution. This localization procedure immediately gives the following expression

$$\begin{aligned} \ell_{\text{loc}i}(p_{i1}, ..., p_{in}, p_i) &= \sum_{j=1}^n w_{ij}(\log p_{ij} + \log p_i) \\ &= \sum_{j=1}^n w_{ij} \log p_{ij} + \log p_i \end{aligned} \tag{2.8}$$

for the i-th log-likelihood contribution. The nonparametric likelihood function for the conditional moment restriction model is[1]

$$\ell(\{\{p_{ij}\}_{j=1}^n\}_{i=1}^n, \{p_i\}_{i=1}^n, \theta) = \sum_{i=1}^n \ell_{\text{loc}i}(p_{i1}, ..., p_{in}, p_i), \qquad \text{subject to (2.5), (2.6).}$$

As in usual models with conditioning, the component $\sum_{i=1}^n \log p_i$ can be dropped from the likelihood function for estimating θ. Note that concentrating out the parameters $\{\{p_{ij}\}_{j=1}^n\}_{i=1}^n$ is easy, since an application of the Lagrange multiplier method as in the previous section yields

$$\begin{aligned}
&\sup_{\{\{p_{ij}\}_{j=1}^n\}_{i=1}^n} \sum_{i=1}^n \sum_{j=1}^n w_{ij} \log p_{ij} \quad \text{subject to (2.5), (2.6)} \\
&\quad = \sum_{i=1}^n \sum_{j=1}^n w_{ij} \log w_{ij} - \sum_{i=1}^n \max_{\gamma_i \in \mathbb{R}^q} \sum_{j=1}^n w_{ij} \log(1 + \gamma_i' g(z_j, \theta)) \\
&\quad := \ell(\theta) + \text{ const.}
\end{aligned}$$

The conditional empirical likelihood (CEL) estimator is the maximizer of the concentrated likelihood:

$$\hat{\theta}_{\text{CEL}} = \operatorname*{argmax}_{\theta \in \Theta} \ell(\theta).$$

Remark 2.3. The CEL estimator achieves the semiparametric efficiency bound of the model (2.4) under weak regularity conditions. While there exist estimators that achieve efficiency in the model, the EL-based estimator has an advantage that finding a preliminary estimator that is consistent is not necessary. A simulation study in Kitamura, Tripathi, and Ahn (2004) indicates that the conditional EL estimator and tests based on it work remarkably well in finite samples. Donald, Imbens, and Newey (2003) propose an alternative estimator for (2.4). Their idea is to use a sequence functions of x as a vector of instruments, then apply EL to the resulting unconditional moment restriction model. By letting the dimension of the instrument vector grow with the sample size in such a way that it spans the "optimal instrument" asymptotically, their procedure also achieves the semiparametric efficiency bound.

Remark 2.4. A topic that is closely related to the above is nonparametric specification testing. Suppose, for example, one is interested in testing the specification of a parametric regression model $E[y|x] = m(x, \theta_0)$, where m is parameterized by a vector $\theta_0 \in \Theta$. The null hypothesis of correct

[1] Kitamura, Tripathi, and Ahn (2004) allow the support of x to be unbounded, and they use trimming to deal with technical problems associated with it. The treatment in this section ignores this issue to simplify presentation. Kitamura, Tripathi, and Ahn (2004) report that the use of trimming factors did not affect the result of their simulation experiments qualitatively.

specification can be written in terms of a conditional moment restriction for the function $g(z,\theta) = y - m(x,\theta), z = (x', y)'$:

$$\textbf{(C)} \qquad E[g(z,\theta)|x] = 0 \text{ for some } \theta \in \Theta.$$

Tripathi and Kitamura (2003) shows that a conditional version of the ELR test applies to the above problem. They propose a simple procedure: reject (**C**) if the maximized value of the conditional empirical likelihood function, which is essentially the one used in Kitamura, Tripathi, and Ahn (2004), is too small. They also calculate the asymptotic power of their test. Their analysis shows that the EL-based testing procedure has an asymptotic optimality property in terms of an average power criterion.

2.3. **Binary Choice Model.** Cosslett (1983) studies an application of the NPMLE procedure to a semiparametric binary choice model. Let $z \in \mathbb{R}^k$ and $\epsilon \in \mathbb{R}$ be a pair of random elements. Assume that ϵ is statistically independent of z and $-\epsilon \sim F_0$. A binary random variable y is drawn according to

$$y = 1\{z'\theta_0 + \epsilon > 0\}, \theta_0 \in \Theta \subset \mathbb{R}^k,$$

conditional on z. The unknown parameters are θ_0 and F_0, and F_0 is treated nonparametrically. The econometrician observes an iid sequence $\{(y_i, z_i)\}_{i=1}^n$, where each (y_i, z_i) is drawn according to the above model.

Cosslett (1983) proposes a semiparametric estimator that applies the NPMLE to estimate F_0. The independence of z and ϵ is important in the construction of Cosslett's NPMLE. If F_0 is known, as in the conventional probit or logit modeling, the log likelihood for θ, conditional on $\{z_i\}_{i=1}^n$, is given by

$$\ell_{\mathrm{P}}(\theta) = \sum_{i=1}^n y_i \log F_0(z_i'\theta) + \sum_{i=1}^n (1-y_i)\log(1 - F_0(z_i'\theta)).$$

Cosslett (1983) suggests to approximate F_0 by an $n-$point discrete distribution. More precisely, consider a distribution with support on the empirical values $\{z_j'\theta\}_{i=1}^n$ and puts probability mass of p_i on each value $z_i'\theta$. This "parameterizes" $F_0(-\epsilon) = \int 1\{\epsilon_i \le -\epsilon\} dF_0$ by $F_n(-\epsilon) = \sum_{j=1}^n 1\{z_j'\theta \le -\epsilon\} p_j$. The corresponding nonparametric log likelihood is obtained by replacing F_0 in ℓ_{P} with F_n:

$$\begin{aligned}\ell(\theta, p_1, ..., p_n) &= \sum_{i=1}^n y_i \log F_n(z_i'\theta) + \sum_{i=1}^n (1-y_i)\log(1 - F_n(z_i'\theta)) \\ &= \sum_{i=1}^n y_i \log\left(\sum_{j=1}^n 1\{z_j'\theta \le z_i'\theta\} p_j\right) + \sum_{i=1}^n (1-y_i)\log\left(1 - \sum_{j=1}^n 1\{z_j'\theta \le z_i'\theta\} p_j\right).\end{aligned}$$

Just like in empirical likelihood, a nested algorithm is typically used for the numerical maximization of $\ell(\theta)$. That is, for a fixed value of θ, maximize $\ell(\theta, p_1, ..., p_n)$ with respect to $\{p_i\}_{i=1}^n$. Let $\ell(\theta)$ denote the maximum value. Cosslett's semiparametric estimator for θ_0 is the maximizer of $\ell(\theta)$ over Θ.

2.4. **Parametric Conditional Model with Endogenous Stratification.** Next, consider semiparametric estimation of models with endogenous stratification, as considered in Cosslett (1993) (see also Cosslett (1981)). Consider a parametric model for the conditional density of a random variable $y \in \mathcal{Y}$ given covariates x

$$f(y|x, \theta_0), \theta_0 \in \Theta \in \mathbb{R}^k.$$

The covariate x obeys a distribution $H_0(x)$, whose density is denoted by $h_0(x)$. H_0 (thus h_0) is treated nonparametrically here, making this model semiparametric. Suppose the data available to the econometrician is subject to endogenous stratification of the following. The sample space $\mathcal{Y}$ is covered by S strata $\{\mathcal{Y}_s\}_{s=1}^S$. From each stratum $\mathcal{Y}_s$, n_s samples $\{((y_{i_s}, z_{i_s})\}_{i_s=1}^{n_s}$ are drawn. The size of total observations is $n = \sum_{s=1}^S n_s$. This is a stratification scheme known as standard stratification. The density of (y_{i_s}, x_{i_s}) is given by

$$\frac{f(y_{i_s}|x_{i_s}, \theta_0)h_0(x_{i_s})}{Q_s(\theta_0, H_0)}, \quad Q_s(\theta, H) = \Pr\{y \in \mathcal{Y}_s\} = \int_{\mathcal{Y}_s} \int f(y|x, \theta) dH dy.$$

Q_s's are often called aggregate shares in the literature. If, contrary to the above statement, the distribution H_0 were known, the model would be completely parametric, with the following log-likelihood function:

$$\ell_{\mathrm{P}}(\theta) = \sum_{s=1}^S \sum_{i=1}^{n_s} \log \frac{f(y_{i_s}|x_{i_s}, \theta)h_0(x_{i_s})}{Q_s(\theta, H_0)}.$$

Note that, unlike the usual conditional model, H_0 (or h_0) does not factor out. One may parameterize H_0 if it is unknown, but its misspecification generally leads to the lack of consistency of the MLE. In practice, specifying the parametric distribution for H_0 is difficult, making the resulting MLE non-robust and unreliable. One way to avoid this is to apply NPMLE to treat H_0 fully nonparametrically. To this end, use an n-dimensional vector

$$(p_{1_1}, ..., p_{n_{11}}, p_{1_2}, ..., p_{n_{22}}, ..., p_{1_S}, ..., p_{n_S S}) \in \Delta$$

to define $H_n(x) = \sum_{s=1}^S \sum_{i_s=1}^{n_s} 1\{x_{i_s} \leq x\} p_{i_s}$, where the distribution $H_n(x)$ has discrete support on the empirical observations $\{\{x_{i_s}\}_{i=1}^{n_s}\}_{s=1}^S$. This yields

$$\hat{Q}_s(\theta) := Q_s(\theta, H_n) = \int_{\mathcal{Y}_s} \sum_{s=1}^S \sum_{i_s=1}^{n_s} f(y|x_{i_s}, \theta) p_{i_s} dy$$

instead of $Q_s(\theta, H_0)$ in the parametric likelihood above. The nonparametric log-likelihood function is therefore

$$\ell(\theta, \{\{p_{i_s}\}_{i=1}^{n_s}\}_{s=1}^{S}) = \sum_{s=1}^{S} \sum_{i_s=1}^{n_s} \log \frac{f(y_{i_s}|x_{i_s}, \theta) p_{i_s}}{\hat{Q}_s(\theta)}.$$

If the parametric model $f(y|x, \theta)$ is a discrete choice model, the above becomes a choice-based sampling model. The value of θ at the maximum of ℓ is the estimator proposed by Cosslett (1981). This estimator is semiparametrically efficient. Cosslett (1993) discusses how to obtain the NPMLE for θ numerically.

2.5. **Conditional Moment Restriction Models with Biased Samples. Data Combination.** Nonparametric likelihood provides a convenient way to combine multiple data sets that are subject to biased-sampling schemes. Qin (1993) notes this fact and considers empirical likelihood-based testing for unconditional moment restriction models. This section focuses on estimation rather than testing, and treats conditional moment restriction models using the technology developed in Kitamura, Tripathi, and Ahn (2004). It is based on the author's joint project with Gautam Tripathi (University of Connecticut).

To fix ideas, consider the following conditional mean regression model,

$$E[y|x] = m(x, \theta_0), \theta_0 \in \Theta \in \mathbb{R}^k, \tag{2.9}$$

though the methodology developed here applies to more general models, such as nonlinear simultaneous equation models or quantile regression models. The vector θ_0 is the parameter of interest. $F_0(y|x)$ and $H_0(x)$ denote the conditional distribution of y given x and the marginal distribution of x, respectively. Let $z = (x_i, y_i)$. Suppose a sample $\{z_i\}_{i=1}^{n_1} = \{(x_i, y_i)\}_{i=1}^{n_1}$ is generated subject to a weighting mechanism with a weighting factor $w(y)$:

$$F_{\text{weighted}}(dz) = \frac{w(y) F_0(dy|x) H_0(dx)}{\int_x \int_y w(y) F_0(dy|x) H_0(dx)}.$$

This occurs when a realization of $z = (x, y)$ is included in the population depending on the value of the endogenous variable y though the function $w(\cdot)$. Various forms for w appear in empirical applications. For example, if y is an $\mathbb{R}_+$-values random variable and $w(y) = y$, this is an example of length biased sampling, which is relevant in duration analysis. If $w(y) = 1\{a \leq y \leq b\}, a, b \in [-\infty, \infty]$, the sample $\{z_i\}_{i=1}^{n_1}$ is generated from a truncated regression model. In the latter case, obviously, the parameter θ is not identified under the conditional mean zero restriction (2.9) from the knowledge of $F_{\text{weighted}}(z)$. The truncated model does not allow point estimation solely based on the sample $\{z_i\}_{i=1}^{n_1}$.

The situation might change, however, if an additional sample is available. The parameter θ may be identified if the property called the "connectedness condition" is satisfied between the two samples (Vardi (1985)). For example, identification is trivially guaranteed if the second sample is a random sample from $F_0(dy|x)H_0(dx)$. Even though identification is not an issue in this case, it is by no means obvious how one should proceed in order to combine the two samples efficiently. An application of nonparametric likelihood offers a natural way to accomplish this challenging task. Let $\{z_i\}_{i=n_1+1}^n = \{(x_i, y_i)\}_{i=n_1+1}^n$ denote the second sample of size n_2. Before obtaining the NPMLE of the semiparametric (2.9) for the semiparametric two sample problem, consider a simple case where the conditional distribution of y given x is parametric and the distribution of x is known for pedagogical purposes. Let $f(y|x;\theta)$ denote the conditional density of y given x parameterized by a vector θ, and $h_0(x)$ the known marginal density of x. Then the parametric log-likelihood function for θ is:

$$\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{n_1} \log\left(\frac{w(y_i)f(y_i|x_i;\theta)h_0(x_i)}{\int_x\int_y w(y)f(y;\theta|x)dydH_0(x)}\right) + \sum_{i=n_1+1}^{n} \log(f(y_i|x_i;\theta)h_0(x_i)). \\
&= \sum_{i=1}^{n_1} \log\left(\frac{w(y_i)f(y_i|x_i;\theta)h_0(x_i)}{\int_x R(x;\theta)dH_0(x)dy}\right) + \sum_{i=n_1+1}^{n} \log(f(y_i|x_i;\theta)h_0(x_i)), \quad R(x;\theta) = \int_y w(y)f(y;\theta|x)dy \\
&= \text{const.} + \sum_{i=1}^{n} \log f(y_i|x_i;\theta) + \sum_{i=1}^{n} \log h_0(x_i) - n_1 \log\left(\int_x R(x;\theta)dH_0(x)\right).
\end{aligned}$$

Note that the density of x does not factor out due to the endogenous biasing factor $w(y)$. This is similar to the problem observed in Section 2.4, where a one sample problem was discussed. If H_0 is unknown, it is possible to estimate it nonparametrically using the NPMLE.

The treatment of the model (2.9), where both $F_0(y|x)$ and $H_0(x)$ are nonparametric functions, is more complicated. If biasing is not present, it is a special case of the conditional moment restriction model in Section 2.2. Therefore the kernel smoothing approach presented there works if the biased sampling mechanism is appropriately incorporated into the nonparametric likelihood function. Define the Nadaraya-Watson weights $\{\{w_{ij}\}_{j=1}^n\}_{i=1}^n$ as before. The discrete distributions $F_n(z|x_i) = \sum_{j=1}^n p_{ij}1\{z_j \le z\}, i = 1, ..., n$ and $H_n(x) = \sum_{i=1}^n p_i 1\{x_i \le x\}$ approximate $F_0(y|x_i)$ and $H_0(x)$. Let $g(z_i, \theta) = y_i - m(x_i, \theta)$. The constraints on the parameter space are

$$\sum_{j=1}^{n} p_{ij} g(z_i, \theta) = 0, i = 1, ..., n, \theta \in \Theta, \tag{2.10}$$

$$\sum_{i=1}^{n} p_{ij} = 1, \quad p_{ij} \ge 0, \quad \sum_{i=1}^{n} p_i = 1, \quad p_i \ge 0 \quad \text{for } 1 \le i, j \le n. \tag{2.11}$$

Define $R_i = \sum_{j=1}^n w(y_j)p_{ij}$. The nonparametric log-likelihood is given by

$$
\begin{aligned}
\ell &= \sum_{i=1}^{n_1}\sum_{j=1}^{n} w_{ij}\log\left(\frac{w(y_i)p_{ij}p_i}{\sum_{i=1}^n\sum_{j=1}^n w(y_j)p_{ij}p_i}\right) + \sum_{i=n_1+1}^{n}\sum_{j=1}^{n} w_{ij}\log(p_{ij}p_i) \\
&= \sum_{i=1}^{n_1}\log\left(\frac{w(y_i)p_{ij}p_i}{\sum_{i=1}^n R_i p_i}\right) + \sum_{i=n_1+1}^{n}\sum_{j=1}^{n}\log(p_{ij}p_i) \\
&= \text{const.} + \sum_{i=1}^{n}\sum_{j=1}^{n}\log p_{ij} + \sum_{i=1}^{n}\log p_i - n_1\log\left(\sum_{i=1}^{n} R_i p_i\right)
\end{aligned}
\tag{2.12}
$$

with the parameter constraints (2.10) and (2.11).

Maximizing (2.12) under (2.10) and (2.11) is harder than the CEL maximization in Section 2.2 for two reasons. First, the parameters $\{p_i\}_{i=1}^n$ for the marginal of x need to be maximized jointly with the other parameters, since they do not factor out from the nonparametric likelihood due to biased sampling. Second, log-likelihood contributions are connected through the term $\log(\sum_{i=1}^n R_i p_i)$. Consequently the duality property that made the calculation of each log-likelihood contribution easy in Section 2.2 does not apply here. Direct numerical maximization of (2.12) over the space of $\{\{p_{ij}\}_{j=1}^n\}_{i=1}^n$, $\{p_i\}_{i=1}^n$ and θ is, of course, unrealistic even for a moderate sample size due to its high dimensionality. Nevertheless, the following algorithm successfully turns the high dimensional problem into a sequence of low dimensional optimization problems.

The idea of the maximization procedure is to combine the Gauss-Seidel algorithm with the duality based algorithm developed for (C)EL. Note that the most difficult part of this maximization problem is the treatment of $\{\{p_{ij}\}_{j=1}^n\}_{i=1}^n$ and $\{p_i\}_{i=1}^n$. Once the concentrated likelihood for fixed values of θ is obtained, the NPMLE for θ is defined as its maximizer. So, keep θ fixed for a moment. As in the standard Gauss-Seidel algorithm, or more precisely the "block Gauss-Seidel algorithm" (Judd (1998)), split the parameters into blocks, so that the optimization of each parameter block, holding the other parameters fixed, is numerically easy to carry out. Then each block is updated sequentially until convergence. Here an appropriate partition is to use $n+1$ blocks: $\{p_i\}_{i=1}^n$, $\{p_{1j}\}_{j=1}^n$,..., $\{p_{nj}\}_{j=1}^n$. The duality-based algorithm reduces the computational cost of maximization within each block substantially. Let $\{\{p_{ij}^{(k)}\}_{j=1}^n\}_{i=1}^n$ and $\{p_i^{(k)}\}_{i=1}^n$ denote the parameter values at the k-th iteration. Similarly, define $R_i^{(k)} = \sum_{j=1}^n w(y_j)p_{ij}^{(k)}$. The $(k+1)$-th iteration of the algorithm consists of the following four steps:

Step 1: This step carries out the maximization over $\{p_i\}_{i=1}^n$, with other parameters fixed at their values in the k-th iteration. It is convenient to introduce an additional parameter $v = \sum_{i=1}^n R_i^{(k)} p_i$. This plays a role similar to a parameter introduced by Qin (1993) ("W" in his notation). Let $\ell^{(k)}(\{p_i\}_{i=1}^n, v)$ denote the nonparametric log-likelihood for $\{p_i\}_{i=1}^n$ (with the other parameters fixed at the "current guess"). The maximum nonparametric likelihood problem to be solved is:

$$\text{maximize } \ell^{(k)}(\{p_i\}_{i=1}^n, v) = \text{const.} + \sum_{i=1}^n \log p_i - n_1 \log v$$
$$\text{subject to } \sum_{i=1}^n p_i = 1, \sum_{i=1}^n R_i^{(k)} p_i = v.$$

An application of the Lagrange multiplier method to the above problem shows that following. The maximizer $\hat{v}$ is defined as the solution to

$$\hat{v} = \sum_{i=1}^n \frac{R_i^{(k)}}{\frac{n_1}{\hat{v}} R_i^{(k)} + n_2},$$

which is a one-dimensional numerical problem. With the value $\hat{v}$, the new guess for p_i's are:

$$p_i^{(k+1)} = \frac{1}{\frac{n_1}{\hat{v}} R_i^{(k)} + n_2}, i = 1, ..., n.$$

Step 2: Set $i = 1$.

Step 3: This step updates $\{p_{ij}\}_{j=1}^n$ (i.e., conditional probabilities at $x = x_i$), which requires some duality result. It maximizes (2.12) with respect to $\{p_{ij}\}_{j=1}^n$ while the other parameters $\{p_i\}_{i=1}^n$, $\{p_{1j}\}_{j=1}^n$,...,$\{p_{(i-1)j}\}_{j=1}^n$, $\{p_{(i+1)j}\}_{j=1}^n$, ..., $\{p_{nj}\}_{j=1}^n$ fixed at $\{p_i^{(k+1)}\}_{i=1}^n$, $\{p_{1j}^{(k+1)}\}_{j=1}^n$,...,$\{p_{(i-1)j}^{(k+1)}\}_{j=1}^n$, $\{p_{(i+1)j}^{(k)}\}_{j=1}^n$, ... ,$\{p_{nj}^{(k)}\}_{j=1}^n$. (It is possible to values from the k-th iteration even for parameters where updated values are available. This corresponds to the block Gauss-Jacobi algorithm.) Define

$$u = \sum_{i'=1}^{i-1} \sum_{j=1}^n w(y_j) p_{i'j}^{(k+1)} + \sum_{j=1}^n w(y_j) p_{ij} + \sum_{i'=i+1}^n \sum_{j=1}^n w(y_j) p_{i'j}^{(k)}$$

as an additional parameter. This corresponds to the parameter v used for updating $\{p_i\}_{i=1}^n$. The optimization problem is:

$$\text{maximize } \ell^{(k)}(\{p_{ij}\}_{j=1}^n, u) = \text{const.} + \sum_{j=1}^n w_{ij} \log p_{ij} - n_1 \log u$$

$$\text{subject to } \sum_{i=1}^n p_{ij} = 1, \sum_{j=1}^n p_{ij} g(z_j, \theta) = 0,$$

$$\sum_{j=1}^n p_{ij} \left(u - w(y_j) - \sum_{i'=1}^{i-1} w(y_j) p_{i'j}^{(k+1)} - \sum_{i'=i+1}^n w(y_j) p_{i'j}^{(k)} \right) = 0.$$

This has a structure identical to the CEL optimization problem discussed in Section 2.2. The roles of $g(z,\theta)$ and θ are played by $[g(z,\theta)', (u - w(y_j) - \sum_{i'=1}^{i-1} w(y_j) p_{i'j} - \sum_{i'=i+1}^n w(y_j) p_{i'j})]'$ and u, respectively. The value of u at the maximum is given by

$$\hat{u} = \operatorname*{argmax}_{u \in \mathbb{R}_+} \ell^{(k)}(u), \quad \ell^{(k)}(u) \tag{2.13}$$

$$= - \max_{\gamma_i \in \mathbb{R}^k, \lambda_i \in \mathbb{R}} \sum_{j=1}^n w_{ij} \log \left(\lambda_i \left[u - w(y_j) - \sum_{i'=1}^{i-1} w(y_j) p_{i'j}^{(k+1)} - \sum_{i'=i+1}^n w(y_j) p_{i'j}^{(k)} \right] + \gamma_i' g(z_j, \theta) \right) - n_1 \log u.$$

Let $\hat{\lambda}_i$ and $\hat{\gamma}_i$ denote the values of λ_i and γ_i that solves (2.13), then the $(k+1)$-th updated values for $\{p_{ij}\}_{j=1}^n$ are:

$$p_{ij}^{(k+1)} = \frac{w_{ij}}{1 + \hat{\lambda}_i \left[u - w(y_j) - \sum_{i'=1}^{i-1} w(y_j) p_{i'j}^{(k+1)} - \sum_{i'=i+1}^n w(y_j) p_{i'j}^{(k)} \right] + \hat{\gamma}_i' g(z_j, \theta)}, j = 1, ..., n.$$

Set $i = i + 1$. Return to the beginning of this step and repeat updating until i reaches n.

Step 4: Update R_i according to $R_i^{(k+1)} = \sum_{j=1}^n w(y_j) p_{ij}^{(k+1)}$. Calculate the updated value of the nonparametric log-likelihood function as follows:

$$\ell^{(k+1)} = \text{const.} + \sum_{i=1}^n \sum_{j=1}^n \log p_{ij}^{(k+1)} + \sum_{i=1}^n \log p_i^{(k+1)} - n_1 \log \left(\sum_{i=1}^n R_i^{(k+1)} p_i^{(k+1)} \right).$$

Repeat Step 1 - Step 4 until convergence. The final value of ℓ obtained in Step 4 is the value of the concentrated log-likelihood at θ, denoted by $\ell(\theta)$. The NPMLE for θ_0 is $\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \ell(\theta)$.

Remark 2.5. If the model is defined unconditionally, the above procedure simplifies. For example, the regression model (2.9) implies the following unconditional moment restriction with a function $d(x)$ of x used as an instrument:

$$E[b(z,\theta)] = 0, \quad \theta \in \Theta, \quad b(z,\theta) = d(x)(y - m(x,\theta)). \tag{2.14}$$

In his important paper, Qin (1993) considered the empirical likelihood ratio testing for the hypothesis $\theta = \theta_0$ with biased samples. In estimating θ_0, the use of the above unconditional moment entails efficiency loss relative to the original conditional moment restriction model (see Tripathi (2005) for discussions on a biased-sample model that is closely related to the one in this remark). It may even cause potential identification failures if $d(\cdot)$ is not chosen appropriately, though in what follows the model is assumed to be identified by the restriction (2.14). Since this formulation does not specify conditioning variables any longer, the unknown nonparametric component of the model is the distribution of z denoted by $F_0(z)$. Define $F_n(z) = \sum_{i=1}^n p_i\{z_i \leq z\}$ (so p_i's parametrize the joint distribution of x and y, not the marginal of x). The nonparametric log-likelihood is

$$\ell = \sum_{i=1}^{n_1} \log \frac{w(y_i)p_i}{\sum_{i=1}^n w(y_i)p_i} + \sum_{i=n_1+1}^{n} \log p_i$$

$$\text{subject to } \sum_{i=1}^n p_i g(z_i, \theta) = 0, \theta \in \Theta, \sum_{i=1}^n p_i = 1, p_i \geq 0.$$

This can be maximized in a similar manner as in the algorithm developed above for the conditional moment restriction model, though the absence of conditioning makes its optimization much simpler. As before, a convenient approach is to define a new parameter $u = \sum_{i=1}^n w(y_i)p_i$ and consider the profile likelihood in which the parameters $\{p_i\}_{i=1}^n$ are concentrated out for fixed θ and u. This amounts to solving the optimization problem:

$$\text{maximize } \sum_{i=1}^n \log p_i - n_1 \log u$$

$$\text{subject to } \sum_{i=1}^n p_i g(z_i, \theta) = 0, \sum_{i=1}^n p_i(u - w(y_j)) = 0, \sum_{i=1}^n p_i = 1, p_i \geq 0.$$

By convex duality, the profile likelihood is easy to compute (see also Qin (1993)):

$$\ell(\theta, u) = \max_{\gamma,\lambda} \sum_{i=1}^n -\log(1 + \gamma' g(z_i, \theta) + \lambda(u - w(y_j))) - n \log u - n \log n.$$

Optimizing $\ell(\theta, u)$ over $\Theta \times \mathbb{R}_+$ yields the NPMLE (or the EL estimator) $\hat{\theta}$ for θ_0. Also, let $\hat{u}$, $\hat{\gamma}$ and $\hat{\lambda}$ denote the values of u, γ and λ at the optimal point, then the NPMLE for $F_0(z)$ is

$$\hat{F}(z) = \sum_{i=1}^n \hat{p}_i\{z_i = z\}, \quad \hat{p}_i = \frac{1}{1 + \hat{\gamma}' g(z_i, \hat{\theta}) + \hat{\lambda}(u - w(y_j))}.$$

Remark 2.6. In Remark 2.5, the distribution F_0 is estimated under the model restriction (2.14). It is sometimes useful to obtain a model-free estimator for F_0. This can be achieved by simply dropping the restriction $\sum_{i=1}^n p_i g(z_i, \theta) = 0, \theta \in \Theta$ in the derivation of $\hat{F}$. The resulting estimator is the celebrated

NPMLE proposed by Vardi (1985), which is, as noted by Pollard (1990), "a far-reaching extension of the classical model for length-biased sampling."

Remark 2.7. The previous remark on Vardi's estimator suggests that it is of interest to drop the conditional moment restriction (2.4) from the NPMLE analysis developed in this section to obtain a model-free estimator for the underlying distribution based on a combination of biased and random samples. Iterate the algorithm Step 1 - Step 4 without the restriction (2.10) until convergence. Let $\{\{\hat{p}_{ij}\}_{j=1}^n\}_{i=1}^n$ denote the final values for p_{ij}'s, then $\hat{F}(z|x_i) = \sum_{j=1}^n p_{ij} 1\{z_j \leq z\}$ is a conditional version of Vardi's NPMLE.

Remark 2.8. Another possible extension of the estimation procedure for (2.4) developed in this section is nonparametric specification testing. For example, one may wish to test the specification the regression model (2.9) nonparametrically by combining possibly biased samples. As noted in Section 2.2, the maximum CEL function provides an effective way of nonparametric specification testing in a random sample. It would be interesting to see whether the same principle remains valid for the model discussed in this section.

3. Alternative Nonparametric Likelihood Procedures

Section 2.1 showed that the NPMLE applied to the moment condition model (2.1) yields Owen's empirical likelihood. Another way to interpret empirical likelihood is to put it in a minimum divergence estimation framework. Interestingly, this approach yields "alternative nonparametric likelihood procedures" that have been proposed in the literature.

Let f and g denote the density functions or the probability functions of distribution functions F and G. Define a "divergence measure" between F and G to be:

$$D(F,G) = \int \phi\left(\frac{f(z)}{g(z)}\right) g(z)dz, \tag{3.1}$$

for a convex function ϕ. It is easy to see that $D(\cdot, G)$ is minimized at G. Let

$$\mathcal{F}(\theta) = \left\{F : \int g(z,\theta)dF = 0, F \text{ is a CDF}\right\}.$$

Then $\mathcal{F} = \cup_{\theta\in\Theta}\mathcal{F}(\theta)$ is the set of all probability distributions that are compatible with the moment restriction (2.1). Now consider the problem of minimizing the divergence $D(F, F_0)$ with respect to $F \in \mathcal{F}$. In other words, a distribution that is "closest" to the true distribution F_0 in the class of distributions $\mathcal{F}$ is sought. Pick a value $\theta \in \Theta$ and define

$$\textbf{(P)} \qquad v(\theta) = \inf_F D(F, F_0) \quad \text{subject to} \int g(z,\theta) f dz = 0, \int f dz = 1.$$

The value $v(\theta)$ is regarded as the minimum divergence between F_0 and the set of distributions that satisfy the moment restriction with respect to $g(z,\theta)$. The nonnegativity of f is maintained if ϕ is modified so that $\phi(z) = \infty$ for $z < 0$; see Borwein and Lewis (1991). The primal problem (**P**) has a dual problem

$$\textbf{(DP)} \qquad v^*(\theta) = \max_{\lambda\in\mathbb{R},\gamma\in\mathbb{R}^q} \left[\lambda - \int \phi^\star(\lambda + \gamma' g(z,\theta))dF_0(z)\right],$$

where $\phi^\star$ is the convex conjugate (or the Legendre transformation) of ϕ, i.e. $\phi^\star(y) = \sup_x [xy - \phi(x)]$. (**DP**) is a finite dimensional unconstrained convex maximization problem. The Fenchel duality theorem implies that $v(\theta) = v^*(\theta)$. Since the true value θ_0 minimizes $v(\theta)$ over Θ, it follows that

$$\theta_0 = \operatorname*{argmin}_{\theta\in\Theta} v^*(\theta). \tag{3.2}$$

Note that the integral in the definition of v^* is the expected value of $\phi^\star(\lambda + \gamma' g(z,\theta))$ with respect the true distribution F_0, which is unknown in practice. A feasible procedure is obtained by replacing the expectation with the sample average, that is

$$\widehat{v}^*(\theta) = \max_{\lambda\in\mathbb{R},\gamma\in\mathbb{R}^q} \left[\lambda - \frac{1}{n}\sum_{i=1}^n \phi^\star(\lambda + \gamma' g(z_i,\theta))\right]. \tag{3.3}$$

Corresponding to (3.2), an appropriate minimum divergence estimator takes the form:

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \widehat{v}^*(\theta).$$

Kitamura (2006b) notes that this is a Generalized Minimum Contrast (GMC) estimator studied by Bickel, Klassen, Ritov, and Wellner (1993). This minimum divergence framework yields empirical likelihood as a special case with $\phi(x) = -\log(x)$ (or equivalently, $\phi^\star(x) = -1 - \log(-y)$). Other choices for ϕ are, of course, possible, and they yield alternative nonparametric likelihood procedures discussed in the literature. For example, $\phi(x) = x\log(x)$ yields the "exponential tilt" estimator (Kitamura and Stutzer (1997)), while $\phi(z) = \frac{1}{2}(x^2 - 1)$ corresponds to the continuous updating GMM estimator (CUE) (Hansen, Heaton, and Yaron (1996)). A convenient parametric family of convex functions known as the Cressie-Read family (Read and Cressie (1988)) subsumes these three important cases. If ϕ belongs to the Cressie-Read family, one can show that the minimum divergence estimator can be written as

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \max_{\gamma\in\mathbb{R}^q} \left[\frac{1}{n}\sum_{i=1}^n \kappa(\gamma' g(z_i,\theta))\right] \tag{3.4}$$

where $\kappa(y) = -\phi^\star(y+1)$. This is essentially equivalent to the Generalized Empirical Likelihood (GEL) estimator by Smith (1997). Smith (2004) provides a detailed account for GEL.

4. Discussion on Efficiency

The previous section introduced the notion of minimum divergence procedures in the context of the moment condition model (2.1). It allows a large class of distance measures, and it is of great interest to explore which distance measure should be used. Moreover, it is possible to consider divergence measures or distances that fall outside of the family (3.1). Obvious examples include the Kolmogorov-Smirnov distance and other CDF-based criteria. Let $\mathbb{F}_n$ denote the empirical distribution function of $\{z_i\}_{i=1}^n$. Using the notation in the previous section, define

$$\theta_n = \underset{\theta\in\Theta}{\operatorname{argmin}} \min_{F\in\mathcal{F}(\theta)} \sup_z |\mathbb{F}_n(z) - F(z)|.$$

This procedure is conceptually reasonable and the resulting minimum Kolmogorov-Smirnov distance estimator would enjoy certain robustness properties. However, it suffers from at least two serious drawbacks. First, the numerical optimization required for this estimator is hard to implement in practice. Second, this estimator does not achieve the semiparametric efficiency bound of the model (2.1) either. Similar comments apply if the Kolmogorov-Smirnov criterion in the above definition is replaced by, say, the Cramer-von Mises criterion. In contrast, the minimum divergence (or GMC) estimators based on the family of divergence measures (3.1) is straightforward to implement due to the dual representation (3.3), and they achieve the semiparametric efficiency bound. In sum, the family (3.1) covers a wide variety of well-known criteria that yield efficient and practical estimators.

The conventional efficiency comparison, however, provides no guide regarding which ϕ-function one should choose in (3.1), since all ϕ's yield equivalent local first order asymptotic approximations for the alternative nonparametric likelihood procedures. This problem calls for a different method of asymptotic efficiency comparison. The Large Deviation Principle (LDP) offers a practical and elegant solution to the problem.

Like the conventional asymptotic method, the LDP provides first order approximations for various estimators and tests. Unlike the conventional theory, which produces local linear approximations, the LDP provides global nonlinear approximations. It is the latter feature that enables the LDP to yield results not obtained by the conventional linear approximations. For example, the LDP shows that EL enjoys many optimality properties that are not shared by, for example, the conventional GMM estimator.

To introduce the concept of the LDP in the context of moment condition models, suppose the econometrician observes iid data $(z_1, ..., z_n)$, where z_i satisfies the restriction (2.1). Let A_n be an event as a result of estimation or testing: for example, if one uses an estimator θ_n to estimate θ_0, one may

consider $A_n = 1\{\|\theta_n - \theta_0\| > c\}$ for a constant c. Then $\Pr\{A_n\}$ is the probability of the estimator missing the true value by a margin larger than c. Or, in testing a null hypothesis H_0, A_n can represent the event that H_0 is accepted. If the null is incorrect, $\Pr\{A_n\}$ is the probability of type II errors. In either way, $\lim_{n\to\infty} \Pr\{A_n\} = 0$ if the estimator or the test is consistent. The LDP also deals with asymptotic properties, but it is concerned with the limit of the form $\lim_{n\to\infty} \frac{1}{n} \log \Pr\{A_n\}$. (If the limit does not exist, one needs to consider $\liminf$ or $\limsup$, depending on the purpose of analysis.) Let $-d \leq 0$ denote the above limit so that $\Pr\{A_n\} \approx e^{-nd}$, which characterizes how fast $\Pr\{A_n\}$ decays. The goal is to obtain a procedure that maximizes the speed of decay d.

Kitamura and Otsu (2005) study the estimation of models of the form (2.1) using the LDP. One complication in the application of the LDP to an estimation problem in general is that an estimator that maximizes the limiting decay rate d with $A_n = 1\{\|\theta_n - \theta_0\| > c\}$ uniformly in unknown parameters does not exist in general, unless the model belongs to the exponential family. A possible way around this issue is to pursue minimax optimality, rather uniform optimality. See Puhalskii and Spokoiny (1998) for a general discussion on such a minimax framework. Note that the probability of the event $A_n = 1\{\|\hat{\theta} - \theta_0\| > c\}$ depends on θ_0 and F_0, therefore the worst case scenario is given by the pair (allowed in the model (2.1)) that maximizes $\Pr\{A_n\}$. Suppose an estimator θ_n minimizes this worst case probability, thereby achieving minimaxity. The limit inferior of the minimax probability provides an asymptotic minimax criterion. Kitamura and Otsu (2005) show that an estimator that attains the lower bound of the asymptotic minimax criterion can be obtained from the EL objective function $\ell(\theta)$ in (2.2) as follows:

$$\hat{\theta}_{\mathrm{ld}} = \operatorname*{argmin}_{\theta\in\Theta} Q_n(\theta), Q_n(\theta) = \sup_{\theta^*\in\Theta:\|\theta^*-\theta\|>c} \ell(\theta^*).$$

Calculating $\hat{\theta}_{\mathrm{ld}}$ in practice is straightforward. If the dimension of θ is high, it is also possible to focus on a low dimensional sub-vector of θ and obtain a large deviation minimax estimator for it, treating the rest as nuisance parameters.

Kitamura (2001) shows that empirical likelihood dominates other methods in terms of the LDP when applied to overidentifying restrictions testing. Applied researchers routinely test overidentifying restrictions of the form

$$\textbf{(O)} \qquad \int g(z,\theta)dF = 0 \quad \text{for some } \theta \in \Theta \text{ and for some distribution function } F,$$

with $\dim(\Theta) = k$ and $g \in \mathbb{R}^q$, $q > k$. The log empirical likelihood under the restriction (**O**) is $\sup_{\theta\in\Theta} \ell(\theta)$; without the restriction, it is $-n \log n$. The ELR test statistic for (**O**) is the difference of the two multiplied by -2. It is asymptotically distributed according to the χ^2 distribution with $q-k$

degrees of freedom under (**O**) (Qin and Lawless (1994)). Using the notation in the previous section, rewrite the above null in an equivalent form: (**O**)' : $F_0 \in \mathcal{F}$. It turns out that ELR for (**O**)' has a property of being uniformly most powerful in an LDP criterion. This optimality property of ELR is formally stated as follows. Generally, a test for (**O**) is represented by a sequence of binary functions $d_n = d_n(z_1, ..., z_n), n = 1, 2, ...$ which takes the value of 0 if the test accept (**O**) and 1 otherwise. The conventional asymptotic power comparison is based on the type II error probabilities of tests that have comparable type I error probabilities. Hoeffding (1963) takes this approach as well, but he evaluates type I and type II errors using LDP. In the present context, size properties of competing tests are made comparable by requiring that, for a parameter $\eta > 0$, each test d_n satisfies

$$\textbf{(L)} \qquad \sup_{P \in \mathcal{P}} \limsup_{n \to \infty} \frac{1}{n} \log P^{\otimes n}\{d_n = 1\} \leq -\eta$$

Therefore η determines the level of a test via large deviations.

Now, use the η in (**L**) to define the ELR test as follows:

$$d_{\mathrm{ELR},n} = \begin{cases} 0 & \text{if } \frac{1}{2n}\mathrm{elr}(\hat{\theta}_{\mathrm{E}L}) \leq -\eta \\ 1 & \text{otherwise.} \end{cases}$$

Kitamura (2001) shows the following two facts under weak regularity conditions.

(I) $d_{\mathrm{ELR},n}$ satisfies the condition (**L**).

(II) For every test d_n that satisfies (**L**),

$$\limsup_{n \to \infty} \frac{1}{n} \log P^{\otimes n}\{d_n = 0\} \geq \limsup_{n \to \infty} \frac{1}{n} \log P^{\otimes n}\{d_{ELR,n} = 0\}$$

for every $P \notin \mathcal{P}$.

Fact (I) shows that the large deviation rate of the type I error probability of the ELR test defined as above satisfies the size requirement (**L**). The left hand side and the right hand side of the inequality in Fact (II) correspond to the LDP of the type II errors of the arbitrary test d_n and that of the ELR test $d_{\mathrm{ELR},n}$, respectively. The two facts therefore mean that, among all the tests that satisfies the LDP level condition (**L**) (and the regularity conditions discussed in Kitamura (2001)), there exists no test that outperforms the ELR test in terms of the large deviation power property. Note that Fact (II) holds for every $P \notin \mathcal{P}$, therefore ELR is uniformly most powerful in terms of LDP. Such a property is sometimes referred to as the Generalized Neyman-Pearson (GNP) optimality.

An alternative way to see why EL works well is to analyze it using higher order asymptotics. Newey and Smith (2004) investigate higher order properties of the GEL family of estimators. To

illustrate their findings, it is instructive to look at the first order condition that the EL estimator satisfies, i.e. $\nabla_\theta \ell(\hat{\theta}_{\mathrm{EL}}) = 0$. A straightforward calculation shows that this condition, using the notation $\hat{D}(\theta) = \sum_{i=1}^n \hat{p}_{\mathrm{EL}i} \nabla_\theta g(z_i, \theta)$ and $\hat{S}(\theta) = \sum_{i=1}^n \hat{p}_{\mathrm{EL}i} g(z_i, \theta) g(z_i, \theta)'$, can be written as:

$$\hat{D}(\hat{\theta}_{\mathrm{EL}})' \hat{S}^{-1}(\hat{\theta}_{\mathrm{EL}}) \bar{g}(\hat{\theta}_{\mathrm{EL}}) = 0; \tag{4.1}$$

see Theorem 2.3 of Newey and Smith (2004). The factor $\hat{D}(\hat{\theta}_{\mathrm{EL}})' \hat{S}^{-1}(\hat{\theta}_{\mathrm{EL}})$ can be interpreted as a feasible version of the optimal weight for the sample moment $\bar{g}(\theta) = \frac{1}{n} \sum_{i=1}^n g(z_i, \theta)$. (4.1) is similar to the first order condition for GMM, though there are important differences. Notice that the Jacobian term D and the variance term S are estimated by $\hat{D}(\hat{\theta}_{\mathrm{EL}})$ and $\hat{S}(\hat{\theta}_{\mathrm{EL}})$ in (4.1). It can be shown that these are semiparametrically efficient estimators of D and S under the moment restriction (2.1). This means that they are asymptotically uncorrelated with $\bar{g}(\theta_0)$, removing the important source of the second order bias of GMM. Moreover, the EL estimator does not involve a preliminary estimator, thereby eliminating another source of the second order bias in GMM. Newey and Smith (2004) formalize this intuition and obtain an important conclusion that the second order bias of the EL estimator is equal to that of the infeasible method-of-moments estimator that optimally weights $\bar{g}$ by the unknown factor $D'S^{-1}$. In contrast, the first order condition of GMM takes a similar form, but the terms that correspond to D and S are inefficiently estimated, causing bias. Newey and Smith (2004) note that the first order conditions of GEL estimators has a form where D is efficiently estimated but S is not, leaving a source of bias that is not present for EL.

Higher oder properties of ELR tests have been studied in the literature as well. One of the significant findings in the early literature of empirical likelihood is the Bartlett correctability of the empirical likelihood ratio test, discovered by DiCiccio, Hall, and Romano (1991). Consider the ELR test statistic for $\mathrm{H}_0 : \theta = \theta_0$ in the model (2.1) with $q = k$. DiCiccio, Hall, and Romano (1991) show that the accuracy of the χ^2 asymptotic approximation for the distribution of the ELR statistic can be improved from the rate n^{-1} to the much faster rate n^{-2} by multiplying it by a factor called the Bartlett coefficient.

5. Conclusion

As demonstrated by the recent developments of empirical likelihood methods in the econometrics literature, nonparametric likelihood offers practical and effective ways to handle empirical problems in economics without making arbitrary assumptions on distributional specifications. The preceding chapters explored various applications of nonparametric likelihood and explored its desirable

properties. Needless to say, there are examples of NPMLE in econometrics that were not discussed discussed above. For instance, the celebrated Cox partial likelihood estimator for a proportional hazard duration model can be interpreted as an NPMLE-based estimator (see van der Vaart (1998)). Cox's model is semiparametric in that the (integrated) baseline hazard function is treated nonparametrically. Consider applying nonparametric likelihood to the integrated baseline hazard using the observed duration values as the support points. Concentrating it out from the likelihood function for a fixed value of the parameter of interest, one obtains the Cox partial likelihood function. The NPMLE also applies to models with censoring and truncation; see Cosslett (1997) and Owen (2001).

Yet another important area not discussed above is the treatment of time series. Suppose the researcher observes a strictly stationary and weakly dependent time series $\{z_1, ..., z_t\}$, and each z_t satisfies the moment condition $E[g(z_t, \theta_0)] = 0, \theta_0 \in \Theta$. Applying EL to this model ignoring dependence is inappropriate; it leads to efficiency loss, and the chi-square asymptotics of the ELR test break down. There are at least three alternative ways to deal with the problem caused by dependence. The first approach is to parametrize the dynamics using a reduced form time series model such as a VAR model (Kitamura (2006b)). While straightforward, this approach involves the risk of misspecifying the dynamics, and reduces the appeal of EL as nonparametric likelihood. The second approach is the blocking method proposed by Kitamura and Stutzer (1997) and Kitamura (1997). The idea is to form data blocks by taking consecutive observations, and apply EL to them. This is termed blockwise empirical likelihood (BEL). BEL preserves the dependence information in the data, in a fully nonparametric manner. See Kitamura and Stutzer (1997) and Kitamura (1997) for details of this approach. The third approach is a hybrid of the first and the second approaches (Kitamura (2006b)). That is, one applies a low order parametric filter to lessen the degree of dependence in the data, then apply BEL to the filtered data. While this does not change the desirable asymptotic property of BEL, it appears to have advantages in finite samples when applied to a time series that is highly persistent.

Section 2.5 presented an example where an application of nonparametric likelihood allows the econometrician to combine multiple data sets in a natural and effective way, while avoiding *ad hoc* parametric assumptions. Other examples of the NPMLE along this line can be found in recent papers on the Ibragimov-Hasminskii model, such as van der Vaart (1994) and Vardi and Zhang (1992). (Bickel, Klassen, Ritov, and Wellner (1993) provides a comprehensive discussion of the Ibragimov-Hasminskii model.) Another subject where the NPMLE is potentially useful is the analysis of measurement error models. Further study of nonparametric likelihood in these applications appears to be fruitful.

References

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

Borwein, J. M., and A. S. Lewis (1991): "Duality relationships for entropy-type minimization problems," *SIAM Journal of Control and Optimization*, 29, 325–338.

Cosslett, S. (1981): "Maximum likelihood estimation for choice-based samples," *Econometrica*, 49, 1289–1316.

——— (1983): "Distribution-free maximum likelihood estimator of the binary choice model," *Econometrica*, 51, 765–782.

——— (1993): "Estimation from endogenously stratified samples," in *Handbook of Statistics*, ed. by G. Maddala, C. Rao, and H. Vinod. Elsevier Science, pp. 1–43.

——— (1997): "Nonparametric maximum likelihood methods," in *Handbook of Statistics*, ed. by G. Maddala, C. Rao, and H. Vinod. Elsevier Science, pp. 385–404.

DiCiccio, T., P. Hall, and J. Romano (1991): "Empirical Likelihood is Bartlett-Correctable," *Annals of Statisics*, 19, 1053–1061.

Donald, S. G., G. W. Imbens, and W. K. Newey (2003): "Empirical Likelihood Estimation and Consistent Tests with Conditional Moment Restrictions," *Journal of Econometrics*, 117, 55–93.

Dvoretzky, A., J. Kiefer, and J. Wolfowitz (1956): "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *Annals of Mathematical Statistics*, 27, 642–669.

Hansen, L. P., J. Heaton, and A. Yaron (1996): "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business and Economic Statistics*, 14, 262–280.

Hoeffding, W. (1963): "Asymptotically Optimal Tests for Multinomial Distributions," *Annals of Mathematical Statistics*, 36, 369–408.

Imbens, G. W. (1997): "One-Step Estimators for Over-Identified Generalized Method of Moments Models," *Review of Economic Studies*, 64, 359–383.

Imbens, G. W., R. H. Spady, and P. Johnson (1998): "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333–357.

Judd, K. L. (1998): *Numerical Methods in Economics*. MIT Press.

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

——— (2001): "Asymptotic optimality of empirical likelihood for testing moment restrictions," *Econometrica*, 69, 1661–1672.

——— (2006a): "Empirical Likelihood," in *New Palgrave Dictionary of Economics*, ed. by S. Durlauf, and L. Blume. Palgrave Macmillan, forthcoming.

——— (2006b): "Empirical Likelihood Methods in Econometrics: Theory and Practice," in *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, ed. by R. Blundell, W. K. Newey, and T. Persson. Cambridge: Cambridge University Press, forthcoming.

Kitamura, Y., and T. Otsu (2005): "Minimax Estimation and Testing for Moment Condition Models via Large Deviations," Manuscript, Department of Economics, Yale University.

Kitamura, Y., and M. Stutzer (1997): "An Information Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65(4), 861–874.

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

Newey, W. K., and R. J. Smith (2004): "Higher order properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–255.

Owen, A. (1988): "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, 75(2), 237–249.

——— (2001): *Empirical Likelihood*. Chapman and Hall/CRC.

Pollard, D. (1990): *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics.

Puhalskii, A., and V. Spokoiny (1998): "On large-deviation efficiency in statistical inference," *Bernoulli*, 4, 203–272.

Qin, J. (1993): "Empirical likelihood in biased sample problems," *Annals of Statistics*, 21, 1182–1196.

Qin, J., and J. Lawless (1994): "Empirical Likelihood and General Estimating Equations," *Annals of Statistics*, 22, 300–325.

Read, T. R. C., and N. A. C. Cressie (1988): *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer.

Ridder, G., and R. Moffitt (2003): "The Econometrics of Data combination," in *Handbook of Econometrics, vol. VI*, ed. by J. Heckman, and E. Leamer. Elsevier Science B.V., forthcoming.

Smith, R. J. (1997): "Alternative semi-parametric likelihood approaches to generalized method of moments estimation," *Economic Journal*, 107, 503–519.

——— (2004): "GEL criteria for moment condition models," Working Paper, University of Warwick.

Tripathi, G. (2005): "Moment based inference with stratified data," Working Paper, Department of Economics, University of Connecticut.

Tripathi, G., and Y. Kitamura (2003): "Testing Conditional Moment Restrictions," *Annals of Statisics*, 31, 2059–2095.

van der Vaart, A. (1994): "Maximum likelihood estimation with partially censored data," *The Annals of Statistics*, 22(4), 1896–1916.

——— (1998): *Asymptotic Statistics*. Cambridge University Press.

Vardi, Y. (1985): "Empirical distributions in selection biased models," *Annals of Statistics*, 13, 178–203.

Vardi, Y., and C.-H. Zhang (1992): "Large sample study of empirical distributions in a random-multiplicative censoring model," *Annals of Statistics*, 20, 1022–1039.