Statistica Sinica **20** (2010), 771-785

INVESTORS' PREFERENCE: ESTIMATING AND DEMIXING OF THE WEIGHT FUNCTION IN SEMIPARAMETRIC MODELS FOR BIASED SAMPLES

Ya'acov Ritov and Wolfgang K. Härdle

The Hebrew University of Jerusalem and Humboldt-Universität zu Berlin

Abstract: We consider a semiparametric model for the weight function in a biased sample model. The object of our interest parametrizes the weight function, and it is non-Euclidean. The model discussed is motivated by the estimation of the mixing distribution of individual utility functions in the DAX market. We discuss the estimation rate of different functionals of the weight functions.

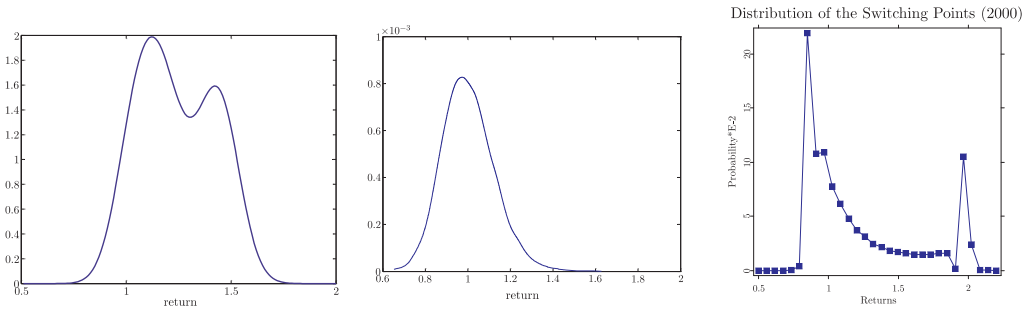
Key words and phrases: Empirical pricing kernel, exponential mixture, inverse problem, mixture distribution, risk aversion.

1. Introduction

A sample $X_1, \ldots, X_n$ is considered biased if it is sampled from a density p which is represented as

$$p(x) = \frac{q(x)w(x)}{\int q(u)w(u)du}. \tag{1.1}$$

Here q is some 'natural' pdf (probability density function) for the problem, representing the 'true' underlying distribution, while w is a given weight function that biases the sample. In a standard example, X represents the severity of the disease, and q is the density of X among patients at admission to the hospital. However, it may be more convenient to take a random sample from the population of patients who are in the hospital at a given time. If the time of hospitalization is proportional to the severity of the case, then the sample is taken from the density p, which is equal to q 'length biased' with $w(x) \equiv x$. Vardi (1985) was the first to systematically analyze these models; asymptotic theory was developed in Gill, Vardi and Wellner (1988); Gilbert, Lele and Vardi (1999) extended the model to the situation where the weight function depends on some parameter, $w(x) = w(x; f)$; the large sample properties were discussed in Gilbert (2000). Equation (1.1) has some similarities to the classical choice-based sample problem, Manski and Lerman (1977), or retrospective case-control studies, Mantel (1973). In fact one can consider the situation as if one has an infinite



(a) (b) (c)

Figure 1. The DAX data, 24/03/2000 half a year look ahead: (a) p, the historical density; (b) q, the risk neutral density; (c) The estimate of f, the mixing density. Figures are taken from DHM.

sample from the control group, and hence q is known, and a finite sample from the control, the biased sample. The likelihood ratio between the two is the given $w(x; f)$. The main difficulty we face in this paper is the particular form of $w(x; f)$ we have.

Technically speaking, our paper is about estimating f, the parameter of the weight function, $w(x) = w(x; f)$. In the model we consider, q is taken as known, while the weight function is parametrized by a non-Euclidean parameter. This brings us to an inverse problem of estimating and demixing the weight function.

In subject matter, our model is motivated by the research on risk aversion and proclivity, and more precisely on the empirical pricing kernel (EPK), see Detlefsen, Härdle and Moro (2007) (hereafter DHM). The EPK describes the apparent utility behavior as function of the individual investors utility function. In this model q is the risk neutral density of asset pricing, and is derived from theoretical considerations. The density p on the other hand is the density of the empirical (historical) prices. See parts (a) and (b) of Figure 1 for an example. In asset pricing the EPK links a risk neutral investor's behavior to individual utilities, which gives in our notation a semiparametric modeling of the weight function w. The integral function of the pricing kernel q/p is the utility function used by a representing individual. Knowing p and q yields the exact form of the utility function, cf. Ait-Sahalia and Lo (2000), and Rosenberg and Engle (2002). The risk neutral (state price) density (SPD) q can be calculated from market data on European options. There are more than 5,000 observations each day for maturity from one week to two years. The SPD can therefore be estimated very precisely. Much empirical research work has demonstrated the so called EPK paradox: the resulting utility function is partially concave and partially convex, more precisely of the Friedman and Savage type, Friedman and Savage (1948).

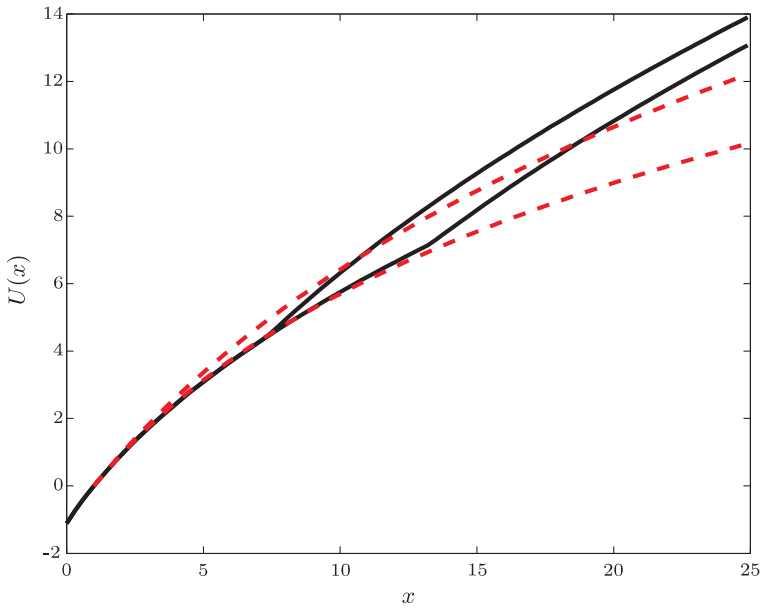


Figure 2. The utility function $U(\cdot;\xi)$ of (3.5) ($\alpha_1 = 2$, $\alpha_2 = 2.25$, $c = 2$) for two different values of ξ (solid lines), and of (3.8) for two values broken lines.

This so called risk aversion puzzle has also been recently discussed in Chabi-Yo, Garcia and Renault (2008); a recursive utility approach to dynamic pricing kernel estimation is published in Gallant and Hong (2007); a fundamental reference on asset pricing theory is the book by Cochrane (2005).

It is assumed in DHM that the observed density of the DAX value has density of the form $p(x) = cq(x)w(x;f)$, where $q \in \{q_\nu, \nu \in N \subseteq \mathbb{R}^d\}$ is the theoretical derived risk neutral density, assumed to follow a given parametric function, and c is a normalization factor, that is, of the type (1.1). The weight function is theoretically derived as

$$w(x;f) = \frac{1}{U'}(x), \tag{1.2}$$

where U is the market utility function, and prime denotes derivative. The market utility is estimated for option data and available historical data, and it also showed the risk aversion puzzle for the DAX stock market. In DHM an aggregation mechanism was proposed that similarly to Chabi-Yo, Garcia and Renault (2008) uses a switching point ξ. This point characterizes the investors switch from a bearish (low return) to a bullish (high return) risk aversion pattern. A graph of two different utility functions $u(\cdot;\xi)$ with switching points $\xi_1 < \xi_2$ is presented in Figure 2.

Simply averaging the utilities is not possible since utilities for different investors are incomparable. One therefore specifies first a utility level u and aggregates the outlooks on the returns R_i with $u = U(R_i; \xi_i)$, $i = 1, 2, \ldots$. The aggregate estimator of the switching return equals average$\{U^{-1}(u, \xi_i), i = 1, 2, \ldots\}$ if all investors have the same market power. Denoting the investors inverse utility function by g and assuming a distribution of switching points, the market utility function U_f is itself assumed to be a function of the mixture of the individual investors:

$$x = U_f^{-1}(u) = \int_{\Xi} g(u; \xi) f(\xi) d\xi. \tag{1.3}$$

Here $\xi \in \Xi$ denotes an investor type, f is the density of the investors' distribution, and $\{g(\cdot; \xi) : \xi \in \Xi\}$ is the (known) class of possible inverse utility functions of the different investors. A subject of type ξ has the inverse utility function $g(\cdot; \xi)$ or, equivalently, he has the utility function $u(\cdot; \xi)$ satisfying $g\{u(x; \xi); \xi\} \equiv x$. The problem we consider is finding the density f. We obtain from (1.1)−(1.3) the representation:

$$p(x) = cq(x) \int \frac{\partial}{\partial u} g(u; \xi) f(\xi) d\xi,$$

where u solves

$$x = \int g(u; \xi) f(\xi) d\xi. \tag{1.4}$$

See Figure 1 for an example taken from DHM of estimates of p, q, and f. See also Figure 2 for an example of $g^{-1}(\cdot; \xi)$.

Aggregation problem (1.3) is a way of aggregating preferences that is not based on the equilibrium theory usually associated with Walras (1874). The situation considered here is of a different type and is hypothetical when applied to real markets. The DAX market data were mentioned as suitable for testing the disaggregation techniques described in the paper.

Aggregation procedure (1.3) relates to the situation where the price of an asset is obtained as the result of a survey of investors (or experts) before they made trades. Thus, this price should be considered as a forecast for the next period, not a reflection of the struggle for limited resources in the market between investors with different preferences and endowments.

The survey proceeds as following. Each market participant is asked what the price will be if the conditions in the market are, for example, extremely good. Extremely good corresponds to some utility level $\tilde{u}_1$ in the minds of investors. In this way all investors agree that they are discussing an economic situation with the same utility level. As the next step, each investor forms his forecast about how high the prices would be in such a situation. Those forecasted prices are recorded and averaged to produce an aggregate opinion of all market participants

(or experts). If the investors have equal market power, their individual opinions will be averaged with equal weights. The forecast for different economic situations corresponding to other utility levels is formed in a similar way.

To sum up, (1.3) describes a mechanism for forming a forecast about future prices. It gives an idea of which opinions prevailed in a group of investors or experts that was able to predict prices correctly before trading, for example if they were more optimistic or pessimistic investors (experts), and to what degree.

In this paper we investigate the estimation of the non-Euclidean parameter f of a few utility functions. The result is typical for inverse problems, in that slightly different assumption yield completely different results. In fact, we present three similar models, similar to those investigated in DHM, that exhibit these behaviors:

(i) there is no consistent estimator of f;

(ii) f can be estimated at a regular nonparametric rate of $n^{-\alpha}$;

(iii) f can be estimated, but at a very slow rate.

Interestingly, there is a a sort of uncertainty principle: the better we can estimate the function $U^{-1}(u)$, the worse we can demix it and estimate f. This is not unexpected. We cannot estimate f well when large differences in f have only minor impact on $\int g(\cdot;\xi)f(\xi)d\xi$.

The structure of the rest of the paper is as follows. In Section 2, we suggest an algorithm for calculating the generalized maximum-likelihood estimator (GMLE) for the semiparametric weight function of the model suggested by DHM. Rates of convergence of the demixing estimator for the DHM's model are discussed in Section 3, as well as of estimates of the mixture itself.

2. EPK: Model and an EM estimator

We consider the EPK problem. We start from (1.4) and we assume that q is known. In practice, it is assumed only to belong to some parametric family $\{q_\nu\}$. However, we deal in the following with rates that are much slower than the parametric $\sqrt{n}$ rate, and the estimate of ν is based on a much larger sample than the estimates of the rest of the parameters. Therefore, the assumption that ν is known considerably simplifies the discussion without impacting the results.

Rewrite (1.4) as

$$p\biggl\{\int g(u;\xi)f(\xi)d\mu(\xi)\biggr\}\int \frac{\partial}{\partial u}g(u;\xi)f(\xi)d\mu(\xi)$$
$$= cq\biggl\{\int g(u;\xi)f(\xi)d\mu(\xi)\biggr\}\biggl\{\int \frac{\partial}{\partial u}g(u;\xi)f(\xi)d\mu(\xi)\biggr\}^2, \tag{2.1}$$

where μ is some dominating measure (e.g., Lebesgue or the counting measure). Noting that the LHS of (2.1) integrates to 1, c can be found to yield

$$p\Big\{\int g(u;\xi)f(\xi)d\mu(\xi)\Big\} = \frac{q\{\int g(u;\xi)f(\xi)d\mu(\xi)\}\int \frac{\partial}{\partial u}g(u;\xi)f(\xi)d\mu(\xi)}{\int q\{\int g(v;\xi)f(\xi)d\mu(\xi)\}\{\int \frac{\partial}{\partial u}g(v;\xi)f(\xi)d\mu(\xi)\}^2 dv}.$$

The market utility $U(x) = U(x;f)$ is given by

$$x \equiv \int g\Big\{U(x;f);\xi\Big\}f(\xi)d\mu(\xi) \equiv \psi_f\Big\{U(x;f)\Big\}.$$

We obtain

$$p(x) = \frac{q(x)\int \frac{\partial}{\partial u}g(U(x;f);\xi)f(\xi)d\mu(\xi)}{\int q(y)\int \frac{\partial}{\partial u}g(U(y;f);\xi)f(\xi)d\mu(\xi)dy} = \frac{q(x)\psi_f'\{\psi_f^{-1}(x)\}}{\int q(y)\psi_f'\{\psi_f^{-1}(y)\}dy}. \tag{2.2}$$

The statistical model assumed by DHM is that we obtain a simple random sample from p, where p is parametrized in (2.2) by the non-Euclidean parameter f. A natural approach is to estimate f by the MLE or a variant of it, which we develop now. Note that $\nabla_f\psi_f(u) = g(u;\cdot)$, and by taking the gradient of $x \equiv \int g\{\psi_f^{-1}(x);\xi\}f(\xi)d\mu(\xi)$ we obtain

$$0 = g\{\psi_f^{-1}(x);\cdot\} + \psi_f'\{\psi_f^{-1}(x)\}\nabla_f\psi_f^{-1}(x).$$

The derivative of the log-likelihood is given therefore by

$$\begin{aligned}\dot\ell_f(\xi) &= \sum_{i=1}^n \frac{1}{\psi_f'\{\psi_f^{-1}(X_i)\}}\left[\frac{\partial}{\partial u}g\{\psi_f^{-1}(X_i);\xi\} - \frac{\psi_f''}{\psi_f'}\{\psi_f^{-1}(X_i)\}g\{\psi_f^{-1}(X_i);\xi\}\right] \\ &\quad - nA_f(\xi), \\ &= \sum_{i=1}^n \frac{1}{\psi_f'\{U_i\}}\left\{\frac{\partial}{\partial u}g\{U_i;\xi\} - \frac{\psi_f''}{\psi_f'}(U_i)g(U_i;\xi)\right\} - nA_f(\xi),\end{aligned}$$

with $U_i = \psi_f^{-1}(X_i)$, and for all $\xi \in \text{suppf}$, where $A_f(\xi)$ is the mean of the first term under f. Since the density of U_i is given by

$$r_f(u) = p\{\psi_f(u)\}\psi_f'(u) = \frac{q\{\psi_f(u)\}\{\psi_f'(u)\}^2}{\int q\{\psi_f(v)\}\{\psi_f'(v)\}^2 dv},$$

we obtain that

$$A_f(\xi) = \frac{\int q\{\psi_f(u)\}\{\psi_f'(u)\frac{\partial}{\partial u}g(u;\xi) - \psi_f''(u)g(u;\xi)\}du}{\int q\{\psi_f(v)\}\{\psi_f'(v)\}^2 dv}.$$

We discusse now how a GMLE can be constructed, and suggest a pseudo-EM algorithm, that is justified as being the limiting result of proper EM algorithms

applied in approximate models. To be clear, the approximation introduced in the following is needed only as a justification for an algorithm applied to the formal model. The algorithm itself is "exact" and maximizes the exact likelihood. The technical problem we want to circumvent is the exact functional dependency of X_i and U_i which affects the EM. As an intermediate step we weaken the functional dependency into a proper statistical dependency.

The model of a random sample from the density p can be well-approximated as $\sigma \to 0$ by a $X_i = \psi_f(U_i) + \varepsilon_i$, $i = 1, \ldots, n$, where $\varepsilon_1, \ldots, \varepsilon_n$ is a random sample from $N(0, \sigma^2)$ independent from the random sample $U_1, \ldots, U_n$ taken from the density r_f. Now, the log-likelihood of the joint density is given by

$$\ell_f = \sum_{i=1}^{n} \Big[\log q\{\psi_f(U_i)\} + 2 \log\{\psi_f'(U_i)\} \Big] - nC_f - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \psi_f(U_i))^2,$$

where $C_f = \log \int q l\{\psi_f(v)\}\{\psi_f'(v)\}^2 dv$. By a well-known formula for the Bayes estimator in the Gaussian measurement error model, here the distribution of $\psi_f(U_i) - X_i$, given X_i, is normal with mean $\sigma^2 f_X'(X_i)/f_X(X_i)$ and second moment $\sigma^4 f_X''(X_i)/f_X(X_i) + \sigma^2$, where f_X is the marginal density of X_i. At the limit as $\sigma^2 \to 0$, the conditional expectation of the log-likelihood, given the X_i's, amounts therefore to replacing U_i by $\psi_f^{-1}(X_i)$. We conclude that the limiting EM algorithm iterates therefore between the following steps.

The E step:

$$U_i \leftarrow \psi_f^{-1}(X_i), \qquad i = 1, \ldots, n, \tag{2.3}$$

The M step:

$$f \leftarrow \operatorname{argmax} \left[\sum_{i=1}^{n} \Big\{ \log q\{\psi_f(U_i)\} + 2 \log\{\psi_f'(U_i)\} \Big\} - nC_f \right].$$

Let $\boldsymbol{U} = (U_1, \ldots, U_n)$, $\boldsymbol{X} = (X_1, \ldots, X_n)$, and denote the E-step by $\boldsymbol{U} = \psi_f^{-1}(\boldsymbol{X})$. The M-step can be accomplished by solving the likelihood equation:

$$0 = \dot{\ell}_f^M(\xi; \boldsymbol{U}) = \sum_{i=1}^{n} \left[\frac{q'\{\psi_f(U_i)\}}{q\{\psi_f(U_i)\}} g(U_i; \xi) + \frac{2}{\psi_f'(U_i)} \frac{\partial}{\partial u} g(U_i, \xi) - \dot{C}_f(\xi) \right], \tag{2.4}$$

for all $\xi \in \operatorname{suppf}$, where

$$\begin{aligned} \dot{C}_f(\xi) &= \frac{\int [(q'\{\psi_f(v)\}/q\{\psi_f(v)\}) g(v; \xi) + (2/\psi_f'(v)) \frac{\partial}{\partial u} g(v, \xi)] q\{\psi_f(v)\}\{\psi_f'(v)\}^2 dv}{\int q\{\psi_f(v)\}\{\psi_f'(v)\}^2 dv} \\ &= \mathrm{E}_f \left[\frac{q'\{\psi_f(U)\}}{q\{\psi_f(U)\}} g(U; \xi) + \frac{2}{\psi_f'(U)} \frac{\partial}{\partial u} g(U, \xi) \right] \\ &= \mathrm{E}_f\{T_f(U; \xi)\}, \quad \text{say.} \end{aligned}$$

However, there is no need in the M-step to find the exact maximizer of the log-likelihood. All that is needed is that the likelihood be strictly increasing (if possible at all) at every M-step. Therefore, the exact M-step given above can be replaced by an approximate M-step, that is obtained by considering an approximate Newton-Raphson solution of (2.4), where the $\mathcal{O}_p(\sqrt{n})$ terms in the Hessian of the log-likelihood are discarded. That is the term

$$\sum_{i=1}^{n} \Big\{ \nabla_f T_f(U_i;\xi) - \mathrm{E}_f \nabla_f T_f(U;\xi) \Big\}.$$

We consider therefore the Newton-Raphson EM (NR-EM) algorithm:

$$f_{i+1} = \begin{cases} \tilde{f}_i \triangleq f_i + H_{f_i}^{-1} \ell_{f_i}^M \{\cdot; \psi_{f_i}^{-1}(\boldsymbol{X})\} & \ell_{\tilde{f}_i} > \ell_{f_i} \\ \text{the solution of (2.3)} & \text{otherwise,} \end{cases}$$

where $H_f : L_2(\mu) \to L_2(\mu)$ is the operator $H_f(\xi,\zeta) = \mathrm{Cov}_f\{T_f(U;\xi), T_f(U;\zeta)\}$.

3. EPK: Rates of Convergence

In the previous section we considered the MLE estimate of f. In this section we consider simple estimators of the type suggested by DHM. Using these estimators we will be able to discuss possible minimax rates of convergence. In essence, we start with a naive nonparametric estimator of the mixture, and in the second step we improve it or demix it for f.

One simple method for demixing the EPK is to start with (1.4) which can be written as

$$1 = c \int \frac{\partial}{\partial u} g(u;\xi) f(\xi) d\xi \frac{q}{p} \Big\{ \int g(u;\xi) f(\xi) d\xi \Big\} = c \frac{\partial}{\partial u} \frac{q}{p} \Big\{ \int g(u;\xi) f(\xi) d\xi \Big\}.$$

Hence $q/p\{\int g(u;\xi) f(\xi) d\xi\} = \alpha + \beta u$ for some α and β, or

$$\int g(u;\xi) f(\xi) d\xi = \Big(\frac{p}{q}\Big)^{-1} (\alpha + \beta u). \tag{3.1}$$

The utility function of an individual is defined up to affine transformation. To assure that it is well defined, we assume that that at the return of 1 the value of the utility is 0, and that of the derivative is 1. In terms of the inverse utility function this translates to $g(0,\xi) \equiv \frac{\partial}{\partial u} g(0,\xi) \equiv 1$. Hence

$$\begin{aligned} \alpha &= \frac{p(1)}{q(1)} \\ \beta &= \frac{p'(1)}{q(1)} - \frac{p(1)}{q(1)} \frac{q'(1)}{q(1)}. \end{aligned} \tag{3.2}$$

The parameter f is therefore the solution of

$$\int g(u;\xi)f(\xi)d\xi = \psi(u) \tag{3.3}$$

for some ψ given explicitly by (3.1) and (3.2). Since q is estimated as a parametric density (based on a much larger sample), and p can be estimated at a standard non-parametric rate based on a direct sample from p, ψ can as well be estimated at a regular density estimation rate.

The analysis of this section starts with (3.3). We assume that ψ and its relevant derivatives can be estimated at a polynomial rate $\|\hat{\psi}^{(i)} - \psi^i\|_\infty = \mathcal{O}_p(n^{-\alpha_i})$ for some $\alpha_i > 0$. The natural estimator suggested by DHM is given by the inverse function of a weighed density estimator. Under strict monotonicity and boundness, the inverse function inherits most properties from the density kernel estimator.

Note that model (3.3) looks like a linear model. For example, if f is approximated by a finite distribution with point mass at $\xi_1, \ldots, \xi_m$, and (3.3) is considered at the k points $u_1, \ldots, u_k$, then it can be written as

$$\hat{\psi}(u_i) = \sum_{j=1}^{m} \beta_j g(u_i;\xi_j) + \varepsilon_i, \qquad i = 1, \ldots, k. \tag{3.4}$$

(3.4) looks like a standard linear model and, indeed, we suggest estimating f by solving it. However, it is not. Most linear model assumptions are violated, e.g., $\varepsilon_1, \ldots, \varepsilon_k$ are not i.i.d. and they are not independent of the random $u_1, \ldots, u_k$.

The basic idea of this section is as follow. We assume that we have some naive nonparametric estimator of ψ. We then proceed to use the pseudo linear model (3.4) to to estimate the mixing distribution and to improve the estimate of ψ itself. We show that this method yields the minimax rates.

How fast can f be estimated? In the rest of the section we present simple examples following DHM. These examples show that in a very similar models very different types of behavior can be obtained. It can be that (i) There is no consistent estimator of f; (ii) f can be estimated at a regular nonparametric rate of $n^{-\alpha}$; (iii) f can be estimated but at a very slow rate. Thus one can suspect that any optimistic result of demixing depends too heavily on assumptions, and are *a priori* not robust (at least in the minimax sense). In particular, any result should be checked to stand against different changes in the model.

3.1. Switching between two utilities

Following DHM assume that for $x, \xi > 0$,

$$U(x;\xi) = \alpha_2(1-c)^{1-1/\alpha_2}\left\{[x-\xi]_+^{1/\alpha_1} \vee (x-c)^{1/\alpha_2}\right\} - \alpha_2(1-c), \tag{3.5}$$

where $\alpha_2 > \alpha_1 > 1$ are given, $c < 0$, and $[x]_+ = x\mathbf{1}(x > 0)$. See Figure 2. Then

$$g(u;\xi) = \min\Big\{\beta^{\alpha_2}\{u + \alpha_2(1-c)\}^{\alpha_2} + c,\ \beta^{\alpha_1}\{u + \alpha_2(1-c)\}^{\alpha_1} + \xi\Big\},$$

where $\beta = \alpha_2^{-1}(1-c)^{-1+1/\alpha_2}$. To simplify the notation and generalize the discussion, we consider a slightly more general case.

Theorem 3.1. *Suppose q is known and bounded away from 0 on a open interval, p has $s > 2$ bounded derivatives, and*

$$g(u;\xi) = \begin{cases} g_2(u) & -\infty < u \le h(\xi) \\ g_1(u) + \xi & \infty > u > h(\xi) \end{cases}, \qquad \xi > 0,$$

where g_1, g_2 are continuous with bounded derivatives, and h given by

$$h^{-1} = g_2 - g_1 \tag{3.6}$$

is a strictly increasing function. Then, f can be estimated with an $\mathcal{O}_p(n^{-(s-2)/(2s+1)})$ error.

Proof. Note that $g(u;\xi)$ is continuous in ξ. Equation (3.3) can be translated to

$$\psi(u) = \int^{h^{-1}(u)} \xi f(\xi)d\xi + g_2(u)F\{h^{-1}(u)\} + g_2(u)\Big\{1 - F\{h^{-1}(u)\}\Big\},$$

where F is the cdf corresponding to the pdf f. Changing variables and considering (3.6),

$$\psi\{h(s)\} = \int^{s} \xi f(\xi)d\xi - sF(s) + g_2\{h(s)\}.$$

Taking a derivative gives $F(s) = h'(s)\{g_2'\{h(s)\} - \psi'\{h(s)\}\}$. Hence estimating F at s is equivalent to the estimation of ψ' at $h(s)$. In other words, $f(\cdot)$ can be estimated at the same rate as the rate of the estimation of second derivative of ψ, which in turn is governed by the rate of estimation of the second derivative of p. Since, by assumption, p has s bounded derivatives, f can be estimated with an $\mathcal{O}_p(n^{-(s-2)/(2s+1)})$ error, cf. Silverman (1986).

3.2. Polynomial and exponential inverse utility function

Theorem 3.1 described a relatively optimistic example. However, modest changes in the inverse utility function may create situations in which f can hardly be estimated, or even not at all.

Here is a pessimistic example:

Theorem 3.2. *Suppose the CRRA (constant relative risk aversion) utility*

$$g(u;\zeta) = (\alpha\zeta^{\alpha-1})^{-1}\Big\{(u+\zeta)^{\alpha} - \zeta^{\alpha}\Big\} + 1, \quad u \in \mathbb{R},\ \zeta \in \mathbb{R}^{+}, \tag{3.7}$$

where α is a known integer. Then there is no consistent estimator of f.

Note that g in (3.7) is scaled such that both its value and its derivative at zero are equal to 1, that is, it represents one branch of (3.5). The proof of Theorem 3.2 is simple. Since α is an integer, $\psi(\cdot)$ is a function of f only through its first α moments. Hence, these moments can be estimated, but no other aspects of f can be estimated or identified.

Seemingly, more and more moments are revealed as $\alpha \to \infty$, and therefore, by the above argument, f is going to be identified at the limit. However, it is not clear that the high moments can be estimated effectively. We consider the limiting case explicitly. The limiting form of the inverse utility function, as $\alpha \to \infty$ and $\alpha/\zeta \to \xi$, is given by

$$g(u;\xi) \equiv \xi^{-1}(e^{u\xi} - 1) + 1. \tag{3.8}$$

The density f is now identified. For example, all its moments can be estimated, e.g., by $\int \xi^{i} f(\xi)d\xi = \psi^{(i+1)}(0)$. We are now going to analyze this model in some detail. We will argue that if $f(\cdot)$ is assumed to have two bounded derivatives, then its value at a point can indeed be estimated, but this can be done only at a very slow convergence rate, slower than any polynomial rate.

Theorem 3.3. *Assume that g is given by* (3.8) *and f is bounded and has two bounded derivatives. Suppose the minimax rate of estimation of ψ is n^{γ}, $\gamma \in (0, 1/2)$. Then there is an estimator $\hat{f}$ such that $\hat{f}(s) - f(s) = \mathcal{O}_p(n^{-\alpha \log\log n/\log n})$ for some α, and for any $\alpha > 0$ there is no estimator $\tilde{f}(s)$ such that $\tilde{f}(s) - f(s) = \mathcal{O}_p(n^{-\alpha/\log\log n})$.*

The proof is given in the on-line supplement, see `http://www.stat.sinica.edu.tw/statistica`.

3.3. Smoothing the empirical estimate and an uncertainty principle

We start, as in the previous subsections, with a nonparametric $\hat{\psi}$. The purpose of this subsection is to show that this initial estimator can be improved considerably by a simple projection.

We argued in Subsection 3.2 that there is no reasonable estimator of f for g given in (3.8). If (3.8) is believed to be true, does this means that there is nothing to do? The surprising answer is no. Although f cannot be estimated per-se, many

of its functionals can be estimated quite easily and quite well. For example, as mentioned in Subsection 3.2, its moments. Similarly $\psi(u)$, another functional of f, can be estimated quite easily, considered as a simple linear functional.

Suppose that f is supported on some compact interval $[a, b]$. Then one can approximate $\psi(u) = \sum_{i=1}^{m} \beta_i u^i + R_m(u)$, where, for some $\tilde{u} \in (0, u)$;

$$0 \le R_m(u) = \frac{1}{(m+1)!}\psi^{m+1}(\tilde{u}) = \frac{1}{(m+1)!}\int_a^b \xi^m e^{\tilde{u}\xi} f(\xi)d\xi \le \frac{b^m e^{ub}}{(m+1)!}. \quad (3.9)$$

Generally speaking, the faster the coefficients β converge to 0, the easier it is to estimate ψ and the harder it is to estimate the mixing density g. As (3.9) shows, we need only a few terms to approximate ψ quite well. In fact we show that in this smooth case, where as on the one hand f can be hardly estimated, ψ can be estimated almost at the parametric rate. This is not an accident — these are two faces of one phenomena. The shape of the observable ψ hardly depends on the fine details of f, and essentially depends only on a few aspects of f. These aspects can be estimated well (and hence ψ can be estimated quite precisely). The other aspects can hardly be estimated and hence f cannot be estimated in a reasonable rate. This yields an uncertainty principle — the more you are certain about ψ the less certain you are about f.

Recall that a function g is called completely monotone if $(-1)^k g^{(k)} \ge 0$, and it is called a Bernstein function if its first derivative is completely monotone. It is well-known (Feller (1966)) that g is completely monotone if, and only if, $g(u) = \int_0^\infty e^{-u\xi} dF(\xi)$. In other words, ψ is a Bernstein function. Nonparametric maximum likelihood estimation for an exponential mixture (and hence completely monotone density) was discussed in Jewell (1982). Balabdaoui and Wellner (2007) discussed the estimation of a k-monotone density.

We assume that there is an estimate $\hat{\psi} = \hat{\psi}_n$ at our disposal. For any $u_1, \ldots, u_k > 0$, let $\Sigma(u_1, \ldots, u_k) \in \mathbb{R}^{k\times k}$, where $\Sigma_{ij}(u_1, \ldots, u_k) = \mathrm{Cov}\{\hat{\psi}(u_i), \hat{\psi}(u_j)\}$. Consider the following assumption:

Assumptions 1. For any n there is $k = k_n$ and $u_1, \ldots, u_k \in (c, d)$, $0 < c < d$, such that the spectral radius of $\Sigma(u_1, \ldots, u_k)$ is $\mathcal{O}(k/n)$, and $\max_i |\mathrm{E}\,\psi(u_i) - \psi(u_i)|^2 = \mathcal{O}(\log n/n)$.

Assumption 1 is satisfied by many nonparametric density and regression estimators, when they strictly under-smooth. We care much more about bias than about variance of the original estimator $\hat{\psi}$. Thus, we have in mind a kernel estimator with bandwidth of order $n^{-1/4+\varepsilon}$. The spectral radius is based on the assumptions that the estimator at points that are a multiple of the bandwidth apart are (almost) independent, for example this is trivially the case with kernel estimators having a compact support. The relationships in the assumption

obtain when the bias of the estimator is $\mathcal{O}(\sigma^2)$, the variance is $\mathcal{O}(1/n\sigma)$, and $k = \mathcal{O}(\sigma^{-1})$.

Consider now the least squares regression of $Y = \{\hat{\psi}(u_1), \dots, \hat{\psi}(u_k)\}^\top$ on the design matrix $Z \in \mathbb{R}^{k\times m}$, $Z_{ij} = u_i^j$. That is, $\hat{\beta} = (Z'Z)^{-1}Z'Y$, where $\hat{\beta} \in \mathbb{R}^m$. Finally let $\tilde{\psi}(u) = \sum_{j=1}^m \hat{\beta}_j u^j$, $u > 0$. We argue that the error achieved by $\tilde{\psi}$ is almost the parametric rate even though $\hat{\beta}$ can be estimated at a strictly lower rate.

Theorem 3.4. *Suppose $g(u;\xi) \equiv \xi^{-1}(e^{u\xi} - 1)$ and that f is supported on a compact interval. Assume 1 holds and $m = m_n = \log n/\log\log n$. Then $k^{-1}\sum_{i=1}^k \{\tilde{\psi}(u_i) - \psi(u_i)\}^2 = \mathcal{O}_p\{(\log n)^2/n\}$.*

Proof. Let β^0 be the true value $\beta_j^0 = \int \xi^{j-1} f(\xi)d\xi/j!$. Write $Y = Z\beta + \varepsilon$, where ε includes both the random error and the bias terms due to both the estimator and the truncation. The latter term is given in (3.9). By standard least squares results,

$$k^{-1}\mathrm{E}\sum_{i=1}^k \Big\{\tilde{\psi}(u_i) - \psi(u_i)\Big\}^2 = k^{-1}\mathrm{E}\Big\{\varepsilon^\top Z(Z^\top Z)^{-1}Z^\top\varepsilon\Big\}$$
$$= k^{-1}\,\mathrm{trace}\Big\{Z(Z^\top Z)^{-1}Z^\top \mathrm{E}\,(\varepsilon\varepsilon^\top)\Big\}.$$

Since $Z(Z^\top Z)^{-1}Z^\top$ is a projection matrix on a m-dimensional space, the RHS is bounded by the largest eigenvalue of $\mathrm{E}\,(\varepsilon\varepsilon^\top)$ times m/k. This has three components (variance and two biases) and hence

$$k^{-1}\mathrm{E}\sum_{i=1}^k \Big\{\tilde{\psi}(u_i) - \psi(u_i)\Big\}^2 = \mathcal{O}\left[\frac{m}{k}\left\{\frac{k}{n} + k\frac{\log n}{n} + k\Big(\frac{b^m}{m!}\Big)^2\right\}\right].$$

The factor k before the last two terms is due to the norm of the unit vector in $\mathbb{R}^k$, and, the last term is by (3.9). The theorem follows by taking $m = \log n/\log\log n$.

A more general result can be based on an assumption like the following.

Assumptions 2. For some c, d and each ε there are $h_{\varepsilon,1}, \dots, h_{\varepsilon,M(\varepsilon)}$ such that

$$\sup_{\xi}\min_{\gamma}\max_{c<u<d}\left|g(u;\xi) - \sum_{j=1}^{M(\varepsilon)}\gamma_j h_j(u)\right| < \varepsilon.$$

Note that clearly the assumption ensures the existence of $\gamma(\cdot)$ such that $\max_{c<u<d}$ $|g(u;\xi) - \sum_{j=1}^{M(\varepsilon)}\gamma_j(\xi)h_j(u)| < \varepsilon$, but then there are also $\beta_j = \int \gamma_j(\xi)f(\xi)d\xi$, $j = 1, \dots, M(\varepsilon)$, such that $\max_{c<u<d}|\psi(u) - \sum_{j=1}^{M(\varepsilon)}\beta_j h_j(u)| < \varepsilon$.

The following theorem can be proved similarly to Theorem 3.4:

Theorem 3.5. *Suppose Assumptions* 1 *and* 2 *hold. Let* $\varepsilon_n = \operatorname{argmin}_\varepsilon\{M(\varepsilon)/n+\varepsilon\}$, *and let* $\tilde{\psi}$ *be the least squares estimate of the regression of* $\hat{\psi}$ *on* $h_{\varepsilon_n,1},\ldots,h_{\varepsilon_n,M(\varepsilon_n)}$. *Then* $k^{-1}\sum_{i=1}^k\{\tilde{\psi}(u_i)-\psi(u_i)\}^2=\mathcal{O}_p(\varepsilon_n)$.

In practice, Theorems 3.4 and 3.5 may seem to be of limited use — a knowledge of the structure of the span of the individual utility functions is needed, and the regression is based on an identified efficient base, which may not be natural. For example, we used a polynomial base for the exponential utility function. The practical approach is a histogram or discrete approximation of f. Does such a procedure yield an effective estimator, an estimator which is both statistically speaking efficient, but at the same time easy to compute and can be be used in off-the-shelf manner?

This is indeed the case. Let $\xi_1,\ldots,\xi_{M(\varepsilon)}$ be reasonably spaced points in the support of f. With the notation introduced after Assumption 2, and by a similar argument, for a vector β on the simplex

$$\sup_u\left|\sum_{j=1}^{M(\varepsilon)}\beta_j g(u;\xi_j)-\sum_{j=1}^{M(\varepsilon)}\beta_j\sum_{l=1}^{M(\varepsilon)}\gamma_l(\xi_j)h_l(u)\right|\le\varepsilon.$$

Hence, one can use the base function $g(\cdot;\xi_1),\ldots,g(\cdot;\xi_{M(\varepsilon)})$ as well.

References

Ait-Sahalia, Y. and Lo, A. (2000). Nonparametric risk-management and implied risk aversion. *J. Econometrics* **94**.

Balabdaoui, F. and Wellner, J. A. (2007). Estimation of a k-monotone density: limit distribution theory and the spline connection. Manuscript.

Chabi-Yo, F., Garcia, R. M. and Renault, R. (2008). State dependence can explain the risk aversion puzzle. *Rev. Finan. Stud.* **21**, 973-1011.

Cochrane, J. H. (2005). *Asset Pricing* (*Revised*). Princeton University Press, Princeton.

Detlefsen, K., Härdle, W. K. and Moro, R. A. (2007). Empirical pricing kernels and investor preferences. SFB649 Discussion paper 2007-017, `http://sfb649.wiwi.hu-berlin.de/fedc/discussionPapers_de.php`.

Feller, W. (1966). *An Introduction to Probability Theory and its Applications, Vol. II.* Wiley, New-York.

Friedman, M. and Savage, L. P. (1948). The utility analysis of choices involving risk. *J. Polit. Economy* **56**, 279-304.

Gallant, A. R. and Hong, H. (2007). A statistical inquiry into the plausibility of Epstein-Zin-Weil Utility. *J. Finan. Econom.* **5**, 523-559.

Gilbert, P. B. (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Ann. Statist.* **28**, 151-194.

Gilbert, P. B., Lele, S. R. and Vardi, Y.(1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika* **86**, 27-43.

Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16**, 1069-1112.

Jewell, N. P. (1982). Mixtures of exponential distributions. *Ann. Statist.* **10**, 479-482.

Manski, C. F. and Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica* **45**, 1977-1988.

Mantel, N. (1973). Synthetic restropective studies and related topics. *Biometrics* **29**, 479-486.

Rosenberg, J. and Engle, R. (2002). Empirical pricing kernels. *J. Finan. Econom.* **64**, 341-372.

Silverman, B., (1986). *Density Estimation*. Chapman and Hall, London.

Vardi, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13**, 178-203.

Walras, M.-E. L. (1874). *Éléments d'économie politique pure, ou théorie de la richesse sociale*.

Department of Statistics, The Hebrew University of Jerusalem 91905, Jerusalem, Israel.

E-mail: yaacov.ritov@gmail.com

CASE - Center for Applied Statistics and Economics, Institute for Statistics and Econometrics, Humboldt-Universität zu, 10178 Berlin, Germany.

E-mail: haerdle@wiwi.hu-berlin.de.

(Received February 2008; accepted February 2009)