Journal of Economic Theory and Econometrics, Vol. 26, No. 4, Dec. 2015, 26–43

A univariate sieve density estimation based on a simulated Kolmogorov-Smirnov test

Hosin Song*

Abstract This paper proposes a simulated Kolmogorov-Smirnov (KS)-based sieve density estimation method. It exploits an objective function which is the difference of two empirical distribution functions, one involved with actual observations and the other with simulated observations. By minimizing the objective function with respect to the sieve parameters, a sieve density/distribution estimator is obtained. The equivalence of the sieve distribution estimator and the true distribution can be tested by the KS test since the KS test statistic is easily obtained from the objective function. The resulting sieve density estimator is shown to be consistent.

Numerical experiments are conducted to verify the performance of the proposed method. Furthermore, the proposed method is applied to estimate the income density in South Korea. Whether the actual observations can be rationalized by the estimated distribution can be tested by the proposed bootstrap test.

Keywords Sieve density/distribution estimation, simulated Kolomogorov-Smirnov test

JEL Classification C14, C15, C18

*Department of Economics, Ewha Womans University, Seoul, Korea, email address: hsong@ewha.ac.kr. I would like to thank two anonymous referees for many helpful comments.

Received March 31, 2015, Revised September 9, 2015, Accepted November 2, 2015

1. INTRODUCTION

A density function is contained in a set of square integrable nonnegative functions whose integration over the entire support is one. The space of density functions is a Hilbert space, and the well-known Parseval's equality implies that any density function can be approximated by the finite number of orthogonal functions arbitrarily well. Sieve density estimation follows from that idea.[1] By increasing the dimension of the sieve space in accordance with the increase of sample size, any density function can be approximated arbitrarily well. In practice, it is well known that the approximation is quite good even with small dimensional sieve space. See Gallant and Nychka (1987), Gabler, Laisney and Lechner (1993), and Bierens and Song (2012).

In the sieve density estimation, mostly the Fourier coefficients-related parameters are not serious considerations because their magnitude is not directly informative to have an idea of the true density function. Instead of these nuisance parameters, we are interested in the overall goodness-of-fit of the sieve density estimator. In this paper, we propose to use the idea of the Kolmogorov-Smirnov (KS) test to evaluate the goodness-of-fit of the sieve estimator, as well as to obtain the sieve estimator. We propose to use the objective function, which is the difference of two empirical distributions where one is involved with actual observations and the other with simulated observations. By minimizing the objective function with respect to the sieve parameters, the sieve density/distribution estimator is obtained. Hence, the equivalence of the sieve distribution function estimator and the true distribution function can be tested via the KS test by using the estimation results. This is very handy since the KS test statistic is directly obtained from the proposed objective function. This paper is similar to Bierens and Song (2012) in that both exploit the difference of actual observations and simulated observations. But, there is a significant difference in that this paper exploits the empirical distribution functions of actual and simulated observations, whereas Bierens and Song (2012) use the empirical characteristic functions. Hence, the proposed estimation method leads to an objective function, which is much less complicated than that of Bierens and Song (2012) since the latter requires some tedious trigonometric calculus. The proposed method is related to the KS test, whereas Bierens and Song (2012) is related to the integrated conditional moment test proposed by Bierens and Ploberger (1997).[2]

In section 2, a simulated KS-based sieve density/distribution estimator is

[1] See Chen (2007) for general sieve estimators and their properties.

[2] The integrated conditional moment test is the Cramer-von Mises type test.

proposed based on the simulated KS-type objective function. Then, the resulting density estimator is shown to be a consistent estimator. In section 3, some numerical experiments are conducted, and the application of the proposed method to the income distribution is illustrated. Section 4 presents some concluding remarks. Proofs and Figures are given in the Appendix. As to notations and symbols, "$\xrightarrow{p}$" and "$\xrightarrow{d}$" denote the convergence in probability and the convergence in distribution, respectively. $\overline{\mathscr{D}}$ denotes the closure of the set $\mathscr{D}$.

2. A SIMULATED KS-BASED SIEVE DENSITY ESTIMATOR

2.1. SIEVE DENSITY REPRESENTATION

Bierens (2008) notices that any absolutely continuous distribution function $F(y)$ can be represented as $F(y) = H(G(y))$ where $H(u)$ is a distribution on the unit interval $[0,1]$, and $G(y)$ is strictly increasing as well as its support contains that of $F(y)$. Hence, any continuous density function $f(y)$ can be represented by a density function $h(u)$ on the unit interval $[0,1]$ via the relationship $f(G^{-1}(u)) = h(u)g(G^{-1}(u))$ where $G(y)$ can be interpreted as the initial guess of the distribution $F(y)$. For details, see Bierens (2008), and Bierens (2014).

Bierens (2008) proposes to approximate $h(u)$ by using orthonormal polynomials on the unit interval. Such approximated density function by the finite number of orthonormal polynomials is called an SNP density function. Throughout this paper, sieve density functions indicate the SNP density functions in Gallant and Nychka (1987). Typical examples of orthonormal polynomials on $[0,1]$ are three-term recurrence series and trigonometric series. Some examples of the three-term recurrence series are: Hermite polynomials, Laguerre polynomials, Legendre polynomials and Chebyshev polynomials. The cosine series, Fourier series and sine series are examples of the trigonometric series. In this paper, we focus on Legendre polynomial based on density in Bierens (2008), and Bierens and Song (2012).[3] Similar to Bierens and Song (2012), the space of density function $h(u)$ is defined as

$$\mathscr{D} = \left\{ h(u) = \frac{(1+\sum_{j=1}^{\infty}\delta_j\rho_j(u))^2}{1+\sum_{j=1}^{\infty}\delta_j^2},\ \sum\nolimits_{j=1}^{\infty}\delta_j^2 \leq \infty \right\}. \tag{1}$$

[3]Of course, the proposed estimation method similarly can be applied to other orthonormal polynomial based representations.

Given an a priori chosen sequence $\overline{\delta}_j > 0$ satisfying $\sum_{j=1}^{\infty} \overline{\delta}_j < \infty$, the space of sieve density functions on the unit interval with SNP order k can be defined as follows.

$$\mathscr{D}_k = \left\{ h_k(u) = \frac{(1+\sum_{j=1}^{k} \delta_j \rho_j(u))^2}{1+\sum_{k=1}^{k} \delta_j^2},\ \sup_{j\geq 1} |\delta_j|/\overline{\delta}_j \leq 1 \right\} \tag{2}$$

where $\rho_k(u)$ is the orthonormal Legendre polynomial satisfying $\rho_0(u) = 1, \rho_1(u) = \sqrt{3}(2u-1)$, and $\frac{\sqrt{k+1}/2}{\sqrt{2k+3}\sqrt{2k+1}}\rho_{k+1}(u) + (0.5-u)\rho_k(u) + \frac{k/2}{\sqrt{2k+1}\sqrt{2k-1}}\rho_{k-1}(u) = 0$ for all $k \in \mathbb{N}$.

2.2. SIMULATED KS-BASED SIEVE ESTIMATOR

Assumption 1. *The SNP order k increases as the sample size n increases, i.e.,* $k \equiv k_n \to \infty$ *as* $n \to \infty$.

Assumption 2. *The true density function* $h_0(u) \in \overline{\cup_{k=1}^{\infty} \mathscr{D}_k} = \mathscr{D}$. *Moreover,* $H_0(u) = \int_0^u h_0(t)dt$, *and* $H_k(u) = \int_0^u h_k(t)dt$ *with* $h_k \in \mathscr{D}_k$.

Assumption 2 is about the idea of sieve estimation. The sieve space $\mathscr{D}_k$ is non-decreasing, and hence $\overline{\cup_{k=1}^{\infty} \mathscr{D}_k} = \mathscr{D}$ since $\cup_{k=1}^{\infty} \mathscr{D}_k$ is dense in $\mathscr{D}$.

Assumption 3. *The true distribution* $F_0(y)$ *and the initial guess* $G(y)$ *are strictly increasing in y. Moreover, the support of G contains that of* F_0.

The true distribution function and density function are denoted by $F_0(y)$ and $f_0(y)$ respectively. Assumptions 1-3 are maintained throughout this paper.

Theorem 1 *Let* $h_k(u) \in \mathscr{D}_k$ *be a sequence of densities of absolutely continuous distributions with* $\lim_{k\to\infty} h_k(u) = h_0(u)$ *for each* $u \in (0,1)$, *and* $H_k(u|\delta_k) = \int_0^u h_k(t)dt$. *If* $h_0(u)$ *is a density function on* $[0,1]$, *then* $\lim_{k\to\infty} \int_0^1 |h_k(u) - h_0(u)|du = 0$.

The proof of Theorem 1 is provided in the Appendix. It is a simple adaptation of Scheffe's lemma.[4]

[4] See Serfling (1980).

Corollary 1 *Theorem 1 implies* $\sup_{u\in[0,1]}|H_k(u)-H_0(u)|\to 0$ *as* $k\to\infty$.

The proof of Corollary 1 is provided in the Appendix.

The proposed simulated KS-based sieve density estimator $\hat{h}_k(u)$ can be obtained by minimizing the objective function $\Omega_n(\delta_k)$ with respect to $\delta_k=(\delta_1,\ldots,\delta_k)$.

$$\hat{h}_k(u)\equiv h(u|\hat{\delta}_k)=\arg\min_{h_k\in\mathscr{D}_k}\Omega_n(\delta_k) \tag{3}$$

where

$$\Omega_n(\delta_k)=\sup_y\left|\frac{1}{n}\sum_{i=1}^n I(Y_i\le y)-\frac{1}{n}\sum_{i=1}^n I(\widetilde{Y}_i\le y)\right| \tag{4}$$

where Y_i is an actual observation which is randomly drawn from the true density function $f_0(y)$, while $\widetilde{Y}_i$ is an independent random drawing from $f_k(y)=h_k(u|\delta_k)g(y)$. Note the determination of δ_k is equivalent to the determination of $h_k(u|\delta_k)$. Given δ_k, $\widetilde{Y}_i$ can be obtained by the accept-reject method using $h_k(u|\delta_k)$.[5] The estimation of δ_k in (3) can be implemented by the simplex method in Nelder and Mead (1965). Once $\hat{h}_k(u)\equiv h(u|\hat{\delta}_k)$ is obtained, then $\hat{f}_k(y)=h_k(G(y)|\hat{\delta}_k)g(y)$ and $\hat{F}_k(y)=H_k(G(y)|\hat{\delta}_k)=\int_0^u h(t|\hat{\delta}_k)dt$ are easily obtained.

The following theorem shows that the proposed simulated KS-based sieve density/distribution estimator minimizing the objective function to be $o_p(1)$ is the consistent estimator.

Theorem 2 *Suppose that* $h_0\in\mathscr{D}$ *with* $f_0(y)=h_0(u)g(y)$ *where the support of* $f_0(y)$ *is contained in the support of* $g(y)$ *and* $u=G(y)$. *Let* $\hat{f}_{k_n}(y)=\hat{h}_{k_n}(u)g(y)$ *where* $\hat{h}_{k_n}(u)$ *is defined in (3). Then, the sieve estimator* $\hat{h}_{k_n}(u)$ *and* $\hat{H}_{k_n}(u)=\int_0^u\hat{h}_{k_n}(t)dt$ *satisfy (i)* $\sup_{u\in[0,1]}|\hat{H}_{k_n}(u)-H_0(u)|\xrightarrow{p}0$ *as* $n\to\infty$, *and (ii)* $\int_0^1|\hat{h}_{k_n}(u)-h(u)|du\xrightarrow{p}0$ *as* $n\to\infty$.

The proof of Theorem 2 is provided in the Appendix.

The main advantage from the objective function (4) is that it enables us to do a goodness-of-fit test of the sieve estimator $\hat{F}_k(y)=H_k(u|\hat{\delta}_k)$ with $y=G^{-1}(u)$

[5]The detailed procedure of the accept-reject method is described in Lemma A in the Appendix. For general accept-reject methods, see Devroye (1986).

by the KS test. Note

$$\sqrt{n}\Omega_n(\hat{\delta}_k) = \sqrt{n}\sup_y \left|\hat{F}_0(y) - \hat{F}_k(y)\right| = \sqrt{n}\sup_{u\in[0,1]} \left|\hat{H}_0(u) - \hat{H}_k(u)\right| \quad (5)$$

where $\hat{F}_0(y) = n^{-1}\sum_{i=1}^n Y_i$ is the empirical distribution of the actual observations, and $\hat{F}_k(y) = n^{-1}\sum_{i=1}^n \widetilde{Y}_i$ is the empirical distribution of simulated observations.

Define the null hypothesis and the alternative hypothesis as follows.

H_0: $\hat{F}_k(y)$ *is the true distribution*

versus

H_1: $\hat{F}_k(y)$ *is not the true distribution.*

If the null hypothesis holds, then

$$\sqrt{n}\Omega_n(\hat{\delta}_k) \xrightarrow{d} \sup_{t\in[0,1]} \mathscr{B}(t),$$

where $\mathscr{B}(t)$ is the Brownian bridge process.[6] Moreover, it is well-known that

$$P\left[\sup_{t\in[0,1]} \mathscr{B}(t) \le d\right] = 1 - 2\sum_{j=1}^{\infty}(-1)^{j+1}\exp(-2j^2d^2).^{7}$$

Instead of $1-2\sum_{j=1}^{\infty}(-1)^{j+1}\exp(-2j^2d^2)$, we can use $1-2\sum_{j=1}^{M}(-1)^{j+1}\exp(-2j^2d^2)$ with sufficiently large M. When $M = 1000$, $P\left[\sup_{t\in[0,1]} \mathscr{B}(t) \le 1.23\right] = 0.90$, $P[\sup_{t\in[0,1]} \mathscr{B}(t) \le 1.36] = 0.95$, and $P[\sup_{t\in[0,1]} \mathscr{B}(t) \le 1.63] = 0.99$. Therefore, if $\sqrt{n}\Omega_n(\hat{\delta}_k)$ is less than 1.23 for sufficiently large n, the null hypothesis $H_0 : \hat{F}_k(y) = F_0(y)$ cannot be rejected at the 10% significance level. Thus, we can test whether the estimated sieve distribution is true or not by the proposed KS goodness-of-fit test.

2.3. TEST OF THE VALIDITY OF $\hat{F}_K(Y)$ USING BOOTSTRAP

The validity of the sieve distribution estimator $\hat{F}_k(y) = \hat{H}_k(G(y))g(y)$ can be verified by testing the null against the alternative hypotheses:

[6]For the detailed proof, see the subsubsection 8.3.1 in Appendix I in Lee (2010).

[7]For details, see Serfling (1980).

H_0: *the actual observations are rationalized by the estimated distribution* $\hat{F}_k(y)$

versus

H_1: *the actual observations are not rationalized by the estimated distribution* $\hat{F}_k(y)$.

Following Bierens and Song (2012), we propose to perform the following bootstrap to test the above null hypothesis.

- For the r_{th} bootstrap replication, independently and randomly draw $U_{r,j}$, $j = 1,2,\ldots,2n$, from the estimator $h(u|\hat{\delta}_k)$, where $u = G(y)$. Then, let $U_{r,j}^{(1)} = U_{r,j}$ for $j = 1,\ldots n$, and $U_{r,j}^{(2)} = U_{r,j}$ for $j = n+1,\ldots,2n$, respectively. Note the superscript 1 and 2 denote the first half and the other half of $U_{r,j}$'s, respectively, and the first subscript r denotes the r_{th} bootstrap replication.

- For each random drawing $U_{r,j}^{(i)}$, compute $\widetilde{Y}_{r,j}^{(i)} = G^{-1}(U_{r,j}^{(i)})$, $i = 1,2$. Then, compute KS_r as follows.

$$KS_r = \sup_y \left| \frac{1}{n}\sum_{j=1}^{n} I(\widetilde{Y}_{r,j}^{(1)} \leq y) - \frac{1}{n}\sum_{i=1}^{n} I(\widetilde{Y}_{r,j}^{(2)} \leq y) \right|,$$

  where $r = 1,2,\ldots,R$, and R is the number of bootstrap replications.

- Find the 90th percentile KS_r, denoted by $\tau_{0.90}$, in the sense that $\Pr[KS_r \leq \tau_{0.90}] = 0.9$. Then compare the original $KS \equiv \sqrt{n}\Omega_n(\hat{\delta}_k)$ with $\tau_{90\%}$.

  If $KS \equiv \sqrt{n}\Omega_n(\hat{\delta}_k) < \tau_{0.90}$, the null hypothesis cannot be rejected at the 10% significance level. That is, the observations are rationalized by the estimated distribution $\hat{F}_k(y)$ at the 10% significance level.

The bootstrap test can play a role of a supplementary test to the goodness-of-fit test addressed in the previous subsection. It verifies the null hypothesis that the actual observations Y_i's are rationalized by the estimated distribution $\hat{F}_k(y) \equiv H(G(y)|\widehat{\delta}_k)$, while the goodness-of-fit test in the previous subsection 2.2 directly tests the null hypothesis $H_0 : \hat{F}(y)$ is true. Fundamentally, both null hypotheses of the two tests are similar. But, the way the tests are done are quite different. The bootstrap test has both parametric and nonparametric properties. It is parametric since $\hat{F}_k(y) = H(G(y)|\hat{\delta}_k)$ is used to generate the bootstrap replications. Moreover, it is nonparametric test since the nonparametric quantiles of

the bootstrap Kolomogorov-Smirnov test statistics are used for the test. It is very likely that both tests lead to the similar test results when the sample size is large.[8]

In the following section, some numerical experiments are conducted to verify the performance of the proposed method and an application of the proposed method is illustrated to estimate the 2012 income distribution in South Korea.

3. MONTE CARLO EXPERIMENTS AND APPLICATION

3.1. NUMERICAL EXPERIMENTS

Suppose a random variable Y follows a Gumbel extreme value distribution with a location parameter μ and a scale parameter σ with the distribution $F_0(y) = \exp(-\exp(-(y-\mu)/\sigma))$.[9] Hence, its probability density function is $f_0(y) = \exp(-(y-\mu)/\sigma)\exp(-\exp(-(y-\mu)/\sigma))/\sigma$. Suppose $\mu = 1$ and $\sigma = 1$. Then, $f_0(y) = \exp(y-1)\exp(-\exp(y-1))$. Obtain n independent random drawings from $f_0(y)$, and treat them as actual observations. Orthonormal Legendre polynomials are used to construct the SNP density function. The initially chosen distribution G is the normal distribution $\mathcal{N}(\overline{Y}, S^2)$ where $S_Y^2 = (n-1)^{-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$. That is, $G(y) = \int_{-\infty}^{y} 1/\sqrt{2\pi S^2}\exp(-(t-\overline{Y})^2/(2S^2))dt$. The proposed simulated KS-based sieve estimation is performed to obtain $\hat{f}(y) = h_k(G(y)|\hat{\delta}_k)$ for three cases of $n = 1000, 5000$ and 10000. In each case, 1000 bootstrap replications are obtained so that the 5th percentile and 95th percentile bootstrap density estimates are obtained. These results are shown in Figures 1-3 in the Appendix. The SNP order is selected by choosing the sieve order k when the objective function cannot decrease any longer by using order $k+1$. Specifically, SNP order k is chosen by following the criterion: choose the smallest k satisfying $\Omega(\widehat{\delta}_k) = \Omega(\widehat{\delta}_{k+1})$ and $\Omega(\widehat{\delta}_m) > \Omega(\widehat{\delta}_{m+1})$ for all m with $1 \leq m \leq k-1$. Note that $\Omega(\widehat{\delta}_k) = \Omega(\widehat{\delta}_{k+1})$ happens when $\widehat{\delta}_{k+1} = (\widehat{\delta}_k', 0)'$. Accordingly, the SNP order is selected to be 3, 5 and 6 respectively.[10] Table 3.1 shows the results of goodness-of-fit of the proposed sieve distribution estimator. Note the selected SNP order increases as the sample size increases.[11] The results suggest that the

[8] This property seems to hold generally in sieve estimation which requires large sample.

[9] This distribution is also called type 1 extreme value distribution.

[10] Throughout this paper, this criterion is used to select the SNP order. The determination of the sieve order can be another issue of interest. See Izenman (1991) to see many suggestions about that issue. Bierens and Song (2012) suggests to use the information criterion to determine the order.

[11] This result reflects the sieve estimation idea in Assumption 1.

Table 1: Goodness-of-fit test results of $\hat{F}_k(u)$

Sample size (n)	1000	5000	10000
Selected sieve order (k_n)	3	5	6
KS test stat. ($\sqrt{n}\Omega_n(\hat{\delta}_k)$)	0.854	0.877	1.130

Note: $P[\sup_{t\in[0,1]}\mathscr{B}(t) \leq 1.23] = 0.90, P[\sup_{t\in[0,1]}\mathscr{B}(t) \leq 1.36] = 0.95$, and $P[\sup_{t\in[0,1]}\mathscr{B}(t) \leq 1.63] = 0.99$.

null hypothesis that $\hat{F}_k(y)$ *is the true distribution* cannot be rejected at the conventional significance level in all three cases. In terms of KS test statistic, the case of $(n,k) = (1000,3)$ is the best among the three cases. The comparison of SNP density estimates and the true density function is shown in Figure 4 in the Appendix.

3.2. AN APPLICATION: INCOME DISTRIBUTION

In this subsection, the proposed simulated KS-based sieve density estimation method is exploited to estimate the density function of the 2012 South Korean household income using household financial survey data by the Bank of Korea, Financial Supervisory Service, and Statistics Korea. The sample size is 9831. The chosen SNP order is three. The estimated density function is shown in Figure 5, and the histogram of actual income and simulated income is shown in Figure 6 in the Appendix.[12] The histogram result suggests that both look similar except on a few intervals.

To justify the validity of the estimated sieve distribution function $\hat{F}_k(y)$, we can test the null hypothesis that *the actual observations are rationalized by* $\hat{F}_k(y)$ against the alternative hypothesis that *the actual observations are not rationalized by* $\hat{F}_k(y)$ via the bootstrap in subsection 2.3. The 90th and 95th percentiles of bootstrap KS statistics are $\tau_{0.90}^{BS} = 1.54$ and $\tau_{0.95}^{BS} = 1.70$ respectively.[13] The KS statistic from the estimation is $\sqrt{n}\Omega_n(\hat{\delta}_3) = 0.52$ which is less than $\tau_{0.90}^{BS} = 1.54$.

[12]The unit of the household income is 10 thousand won.

[13]The histogram of the bootstrap KS statistics is provided in Figure 7 in the Appendix for readers who may be interested in it.

Hence, the null hypothesis is not rejected at the 10% significance level. Therefore, the actual observations can be rationalized by the estimated distribution $\hat{F}_k(y)$. Note that the null hypothesis: $\hat{F}_k(y)$ *is the true distribution* is also accepted at 5% and 10% significance level respectively when applying the goodness-of-fit test to this application.[14]

4. CONCLUDING REMARKS

In this paper, we propose a simulated sieve density estimation method based on orthonormal Legendre polynomials. In particular, we propose to use the simulated KS-type objective function, which is the difference of two empirical distribution functions: one is involved with actual observations and the other with simulated observations. By minimizing the objective function with respect to the sieve parameters, a sieve density estimator is obtained. The sieve distribution estimator automatically follows from the density estimator.

The main advantage of the proposed objective function lies in the ease in constructing the objective function, which is contrast to the complicated Cramer-von Mises type objective function such as the integrated moments in Bierens and Song (2012). The final objective function value can be used to evaluate the goodness-of-fit of the proposed estimator. Particularly, the equivalence of the distribution estimator and the true distribution can be tested by the KS test since the KS test statistic is easily obtained from the estimation results. Moreover, the validity of the sieve estimator can be verified by the proposed bootstrap test by testing whether the actual observations can be rationalized by the estimated distribution.

Some numerical experiments are conducted to confirm the performance of the proposed estimation method and the KS test. Moreover, as a supplementary test to the KS test, the proposed bootstrap test is applied to the household income distribution in South Korea.

[14]This is the expected result. Note that the critical values for the KS test are $\tau_{0.90} = 1.23$ and $\tau_{0.95} = 1.36$ respectively.

REFERENCES

Bierens, H. J. (2008), "Semi-Nonparametric Interval-Censored Mixed Proportional Hazard Models: Identification and Consistency Results ", *Econometric Theory*, 24, 749-794

Bierens, H. J. (2014), "The Hilbert Space Theoretical Foundation of Semi-Nonparametric Modeling", Chapter 1 in *Applied Nonparametric and Semiparametric Econometrics and Statistics*, eds. Jeffrey S. Rcine, Liangjun Su and Aman Ullah, Oxford University Press

Birens, H. J. and W. Ploberger (1997) "Asymptotic Theory of Integrated Conditional Moment Tests", *Econometrica*, 65, 1129-1151

Birens, H. J. and H. Song (2012) "Semi-nonparametric estimation of independently and identically repeated first-price auctions via an integrated simulated moments method ", *Journal of Econometrics*, 168, 108-119

Chen, X. (2007), "Large Sample Sieve Estimation of Semi-nonparametric Models", Chapter 76 in *Handbook of Econometrics*, Vol. 6B, eds. James J. Heckmann and Edward E. Leamer, Amsterdam: North-Holland

Devroye, L. (1986), "Non-Uniform Random Variate Generation", Springer-Verlag

Gabler, S., F. Laisney and M. Lechner (1993), "Seminonparametric Estimation of Binary-Choice Models With an Application to Labor-Force Participation", *Journal of Business & Economic Statistics*, 11, 61-80

Gallant, A. R. and D. W. Nychka (1987), "Semi-Nonparametric Maximum Likelihood Estimation", *Econometrica*, 55, 363-390

Izenman, A. J. (1991), "Review Papers: Recent Developments in Nonparametric Density Estimation", *Journal of the American Statistical Association* , 86, 205-224

Lee, M. (2010), "Micro-Econometrics: Methods of Moments and Limited Dependent Variables", *Springer*

Nelder, J.A. and R. Mead (1965), " A simplex method for function minimization", *Computer Journal*, 7, 308-313

Serfling, R. J. (1980), "Approximation Theorems of Mathematical Statistics", *John Wiley and Sons*

Song, H. (2007), "Nonparametric Identification and Semi-nonparametric Estimation of First-Price Auctions", Ph.D. dissertation, *The Pennsylvania State University*

A. APPENDIX

Proof of Theorem 1. Let $I(\cdot)$ be the indicator function. Then,

$$\begin{aligned}|h_k(u)-h_0(u)| &= (h_k(u)-h_0(u))I(h_k(u)\geq h_0(u))+(h_0(u)-h_k(u))I(h_k(u)<h_0(u))\\ &= (h_k(u)-h_0(u))[1-I(h_k(u)<h_0(u))]+(h_0(u)-h_k(u))I(h_k(u)<h_0(u))\\ &= (h_k(u)-h_0(u))+2(h_0(u)-h_k(u))I(h_k(u)<h_0(u)).\end{aligned}$$

Hence,

$$\begin{aligned}\int_0^1 |h_k(u)-h_0(u)|du &= \int_0^1 (h_k(u)-h_0(u))du+2\int_0^1 (h_0(u)-h_k(u))I(h_k(u)<h_0(u))du\\ &= \int_0^1 h_k(u)du-\int_0^1 h_0(u)du\\ &\quad +2\int_0^1 (h_0(u)-h_k(u))I(h_k(u)<h_0(u))du\to 0\end{aligned}$$

since $\int_0^1 h_k(u)du=\int_0^1 h_0(u)du=1$, and $\int_0^1 (h_0(u)-h_k(u))I(h_k(u)<h_0(u))du\to 0$ by the dominated convergence theorem since $(h_0(u)-h_k(u))I(h_k(u)<h_0(u)))<h_0(u)<M$ for some $M>0$, and $\lim_{k\to\infty}(h_0(u)-h_k(u))=0$. *Q.E.D.*

Proof of Corollary 1. It is noted that

$$\begin{aligned}\sup_{u\in[0,1]}|H_k(u)-H_0(u)| &= \sup_{u\in[0,1]}\left|\int_0^u (h_k(t)-h_0(t))dt\right|\\ &\leq \sup_{u\in[0,1]}\int_0^u |h_k(t)-h_0(t)|\,dt=\int_0^1 |h_k(t)-h_0(t)|\,dt.\end{aligned}$$

Therefore, $\int_0^1 |h_k(t)-h_0(t)|\,dt\to 0$ implies $\sup_u |H_k(u)-H_0(u)|\to 0$ as $k\to\infty$. *Q.E.D.*

Accept-reject method. The following lemma from Song (2007) is used to implement the accept-reject method in this paper.

Lemma A. *Let $f_k(\cdot)$ be a density function from which we want to draw a random variable Y, and let $g(\cdot)$ be a density function from which it is easy to draw a random variable Y_0. The proposed accept-reject method below (steps $1-4$) generates Y.*
Step 1: *Find a constant $\overline{c}\geq 1$ such that $f_k(y)\leq \overline{c}g(y)$ for all y.*
Step 2: *Draw an Y_0 from $g(y)$.*
Step 3: *Draw a U from the uniform distribution on* $[0,1]$.

Step 4: *If* $U \leq \overline{c}^{-1} f_k(Y_0)/g(Y_0)$ *then set* $Y = Y_0$, *else redo steps* $2-4$.[15]

Note $f_k(y) = h(G(y)|\delta_k)g(y)$ and the value $\overline{c}$ is determined by a grid search $\overline{c} = \sup_{0\leq u\leq 1} h(u|\delta_k)$, and step 2 can be done by setting $Y_0 = G^{-1}(U_0)$, where U_0 is a random drawing from the uniform $[0,1]$ distribution. The uniform random variable U in step 3 has to be drawn independently of U_0, so that Y_0 and U are independent. Then step 4 yields a random drawing Y from the density function $f_k(y)$.

Proof Theorem 2.

Part (i). We need to show $\sup_y \left| \frac{1}{n}\sum_{i=1}^n I(\widetilde{Y}_{r,i} \leq y) - F_0(y) \right| \xrightarrow{p} 0$. First, it is noted that

$$\begin{aligned}
&\Omega_n(\delta_{k_n}) \\
&= \sup_y \left| \frac{1}{n}\sum_{i=1}^n I(Y_i \leq y) - \frac{1}{n}\sum_{i=1}^n I(\widetilde{Y}_{r,i} \leq y) \right| \\
&= \sup_y \left| \left(\frac{1}{n}\sum_{i=1}^n I(Y_i \leq y) - F_0(y) \right) - \left(\frac{1}{n}\sum_{i=1}^n I(\widetilde{Y}_{r,i} \leq y) - F_0(y) \right) \right| \\
&\leq \sup_y \left| \frac{1}{n}\sum_{i=1}^n I(Y_i \leq y) - F_0(y) \right| + \sup_y \left| \frac{1}{n}\sum_{i=1}^n I(\widetilde{Y}_{r,i} \leq y) - F_0(y) \right|,
\end{aligned} \tag{6}$$

and that

$$\begin{aligned}
&\Omega_n(\delta_{k_n}) \\
&= \sup_y \left| \frac{1}{n}\sum_{i=1}^n I(Y_i \leq y) - \frac{1}{n}\sum_{i=1}^n I(\widetilde{Y}_{r,i} \leq y) \right| \\
&= \sup_y \left| \left(\frac{1}{n}\sum_{i=1}^n I(Y_i \leq y) - F_0(y) \right) - \left(\frac{1}{n}\sum_{i=1}^n I(\widetilde{Y}_{r,i} \leq y) - F_0(y) \right) \right| \\
&= \sup_y \left| \left(\frac{1}{n}\sum_{i=1}^n I(\widetilde{Y}_{r,i} \leq y) - F_0(y) \right) - \left(\frac{1}{n}\sum_{i=1}^n I(Y_i \leq y) - F_0(y) \right) \right| \\
&\geq \sup_y \left| \frac{1}{n}\sum_{i=1}^n I(\widetilde{Y}_{r,i} \leq y) - F_0(y) \right| - \sup_y \left| \frac{1}{n}\sum_{i=1}^n I(Y_i \leq y) - F_0(y) \right|.
\end{aligned} \tag{7}$$

[15] It is important to restart from step 2, because Y_0 and U need to be mutually independent.

(6) and (7) is reduced to the following:

$$\begin{aligned}
&\sup_y \left| \frac{1}{n} \sum_{i=1}^n I(\widetilde{Y}_{r,i} \le y) - F_0(y) \right| - \sup_y \left| \frac{1}{n} \sum_{i=1}^n I(Y_i \le y) - F_0(y) \right| \\
&\le \Omega_n(\delta_{k_n}) \\
&\le \sup_y \left| \frac{1}{n} \sum_{i=1}^n I(Y_i \le y) - F_0(y) \right| + \sup_y \left| \frac{1}{n} \sum_{i=1}^n I(\widetilde{Y}_{r,i} \le y) - F_0(y) \right|.
\end{aligned} \tag{8}$$

Glivenko-Cantelli theorem implies that

$$\sup_y \left| \frac{1}{n} \sum_{i=1}^n I(Y_i \le y) - F_0(y) \right| \xrightarrow{p} 0. \tag{9}$$

Therefore, it follows from (8) and (9) that

$$\Omega_n(\delta_{k_n}) \xrightarrow{p} 0 \text{ if and only if } \sup_y \left| \frac{1}{n} \sum_{i=1}^n I(\widetilde{Y}_{r,i} \le y) - F_0(y) \right| \xrightarrow{p} 0.$$

Note that $\sup_y \left| \frac{1}{n} \sum_{i=1}^n I(\widetilde{Y}_{r,i} \le y) - F_0(y) \right| \xrightarrow{p} 0$ implies $\sup_y \left| \hat{F}_{k_n}(y) - F_0(y) \right| \xrightarrow{p} 0$, which is equivalent to $\sup_u \left| \hat{H}_{k_n}(u) - H_0(u) \right| \xrightarrow{p} 0$ as $n \to \infty$. *Q.E.D.*

Part (ii). Similar to the part (i), $|\hat{h}_{k_n}(u) - h_0(u)| = |\hat{h}_{k_n}(u) - h_{k_n}(u) + h_{k_n}(u) - h_0(u)|$ has the following inequality.

$$\begin{aligned}
&|\hat{h}_k(u) - h_k(u)| - |h_k(u) - h_0(u)| \\
&\qquad \le |\hat{h}_k(u) - h_0(u)| \le |\hat{h}_k(u) - h_k(u)| + |h_k(u) - h_0(u)|
\end{aligned} \tag{10}$$

where n is suppressed in k_n for convenience. Hence,

$$\begin{aligned}
&||\hat{h}_k(u) - h_k(u)||_1 - ||h_k(u) - h_0(u)||_1 \\
&\qquad \le ||\hat{h}_k(u) - h_0(u)||_1 \le ||\hat{h}_k(u) - h_k(u)||_1 + ||h_k(u) - h_0(u)||_1
\end{aligned} \tag{11}$$

where $||h(u)||_1 \equiv \int_0^1 |h(u)|du$. Note $||h_k(u) - h_0(u)||_1 \to 0$ by Theorem 1. Hence, (11) becomes the following inequality

$$\begin{aligned}
&||\hat{h}_k(u) - h_k(u)||_1 - o(1) \\
&\qquad \le ||\hat{h}_k(u) - h_0(u)||_1 \le ||\hat{h}_k(u) - h_k(u)||_1 + o(1).
\end{aligned} \tag{12}$$

For given $k \in \mathbb{N}$, $||\hat{h}_k(u) - h_k(u)||_1 \xrightarrow{p} 0$. Therefore, $||\hat{h}_k(u) - h_0(u)||_1 \xrightarrow{p} 0$ as $n \to \infty$. *Q.E.D.*

Figure 1: SNP density estimate when SNP order is 3 and $n = 1000$

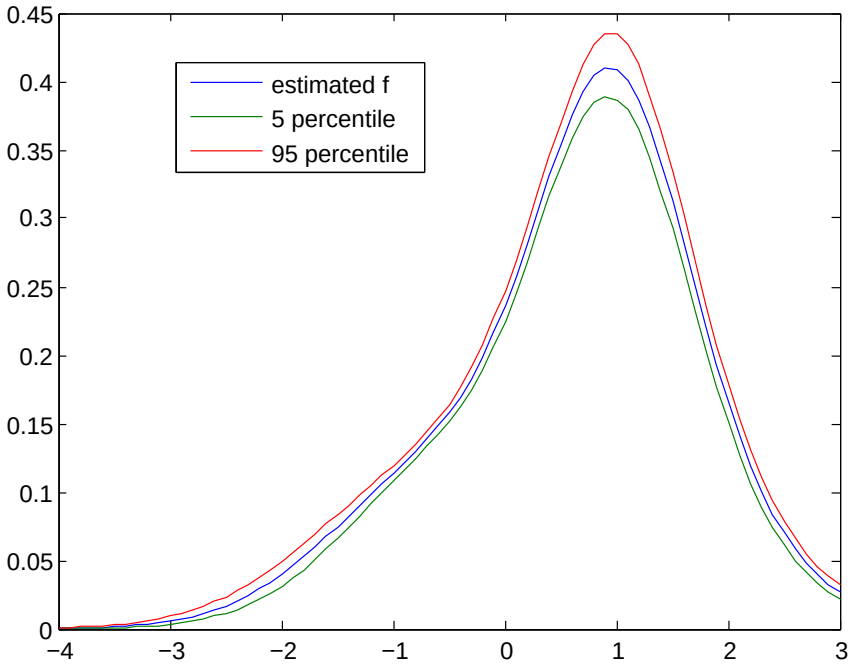

Figure 2: SNP density estimate when SNP order is 5 and $n = 5000$

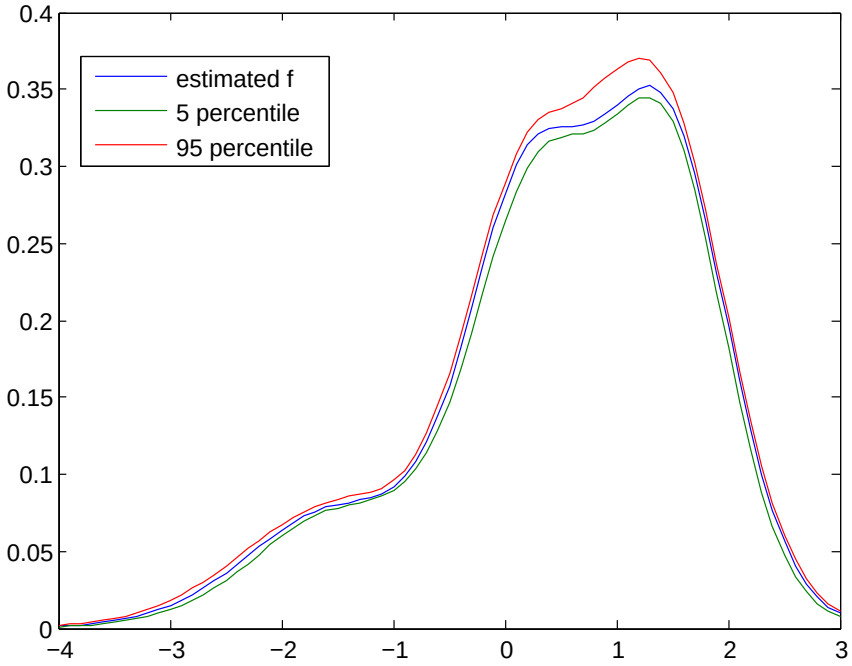

Figure 3: SNP density estimate when SNP order is 6 and $n = 10000$

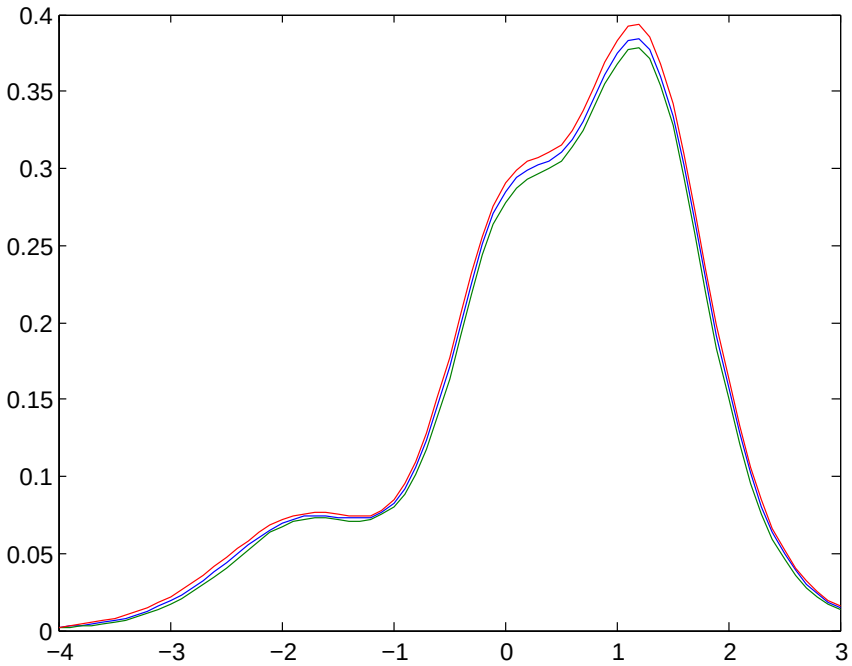

Figure 4: Comparison of SNP density estimates when $n = 1000, 5000, 10000$

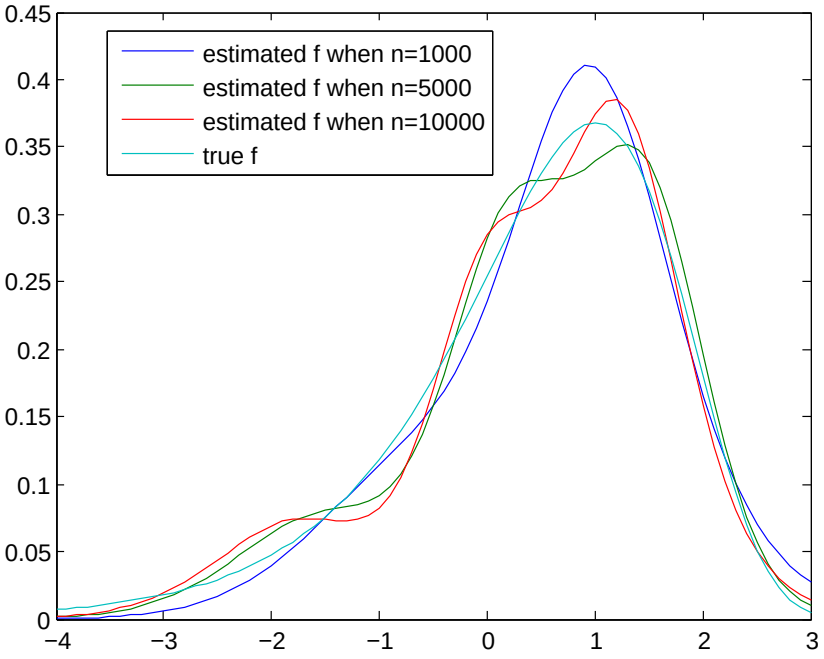

Figure 5: Estimated density function of the household income with SNP order 3

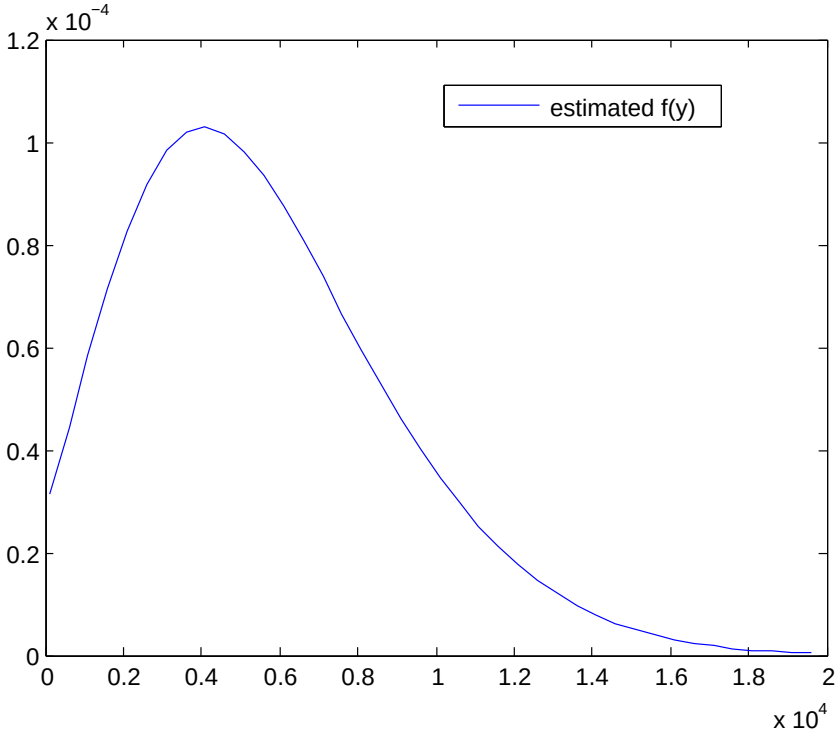

Figure 6: Histogram of actual income and simulated income

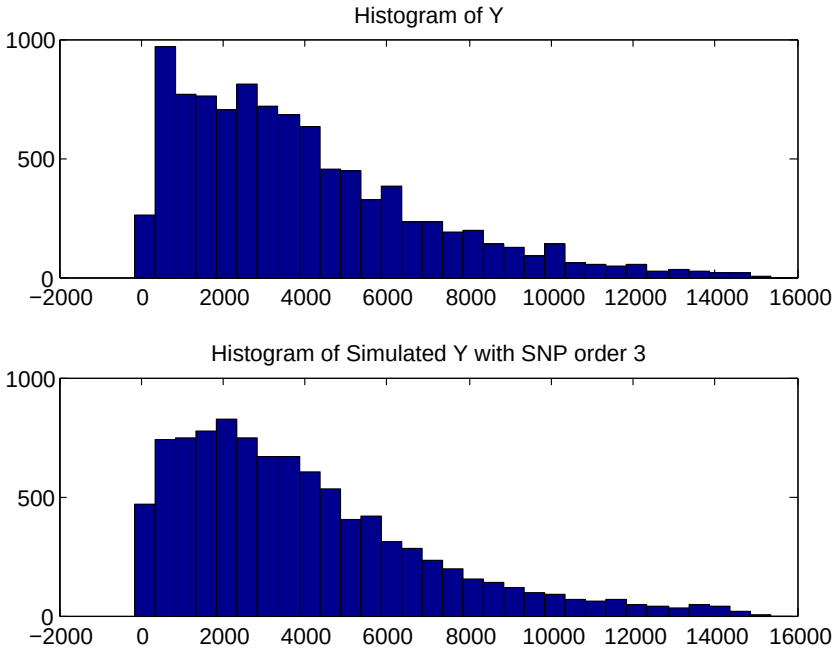

Figure 7: Histogram of bootstrap KS test statistic with $\hat{\delta}_3$