Information-Theoretic Deconvolution Approximation of Treatment Effect Distribution

Ximing Wu* & Jeffrey M. Perloff†

Abstract

This study proposes an information-theoretic deconvolution method to approximate the entire distribution of individual treatment effect. This method uses higher-order information implied by the standard average treatment effect estimator to approximate the underlying distribution of individual treatment effect using the method of maximum entropy density. With minimum assumptions, this method is able to approximate the treatment effect distribution even if it is entirely random or dependent on unobservable covariates. The asymptotic properties of the proposed estimator is discussed. This estimator minimizes the Kullback-Leibler distance between the underlying density and the approximations. Monte Carlo simulations and experiments with real data demonstrate the efficacy and flexibility of the proposed deconvolution estimator. This method is applied to data from the U.S. Job Training Partnership ACT (JTPA) program to estimate the distribution of program impacts on individual earnings.

*Department of Agricultural Economics, Texas A&M University; email: xwu@ag.tamu.edu

†Department of Agricultural and Resource Economics, University of California at Berkeley; email: perloff@are.berkeley.edu

We want to thank Jeffrey Smith for providing the JTPA data and helpful suggestions. We thank participants at University of Toronto seminar, the Midwest Econometrics Group 2005 meeting and the Econometrics Society 2006 winter meeting, especially Tong Li and Kevin Song, for valuable comments and suggestions. The first author acknowledges financial support from the Social Science and Human Research Council of Canada.

1 Introduction

The past two decades have seen a surge in econometric work on evaluating program effects. Prior to the last few years, the literature has focused on the average treatment effect, which is the first moment of the treatment effect distribution. However, the average treatment effect may mask important distributional changes. Recently, researchers have developed new methods to estimate heterogeneous treatment effects and the distribution of treatment outcome. In this study, we present a novel estimator that approximates the entire individual treatment effect distribution based on the standard average treatment effect model. The proposed estimator has an appealing information-theoretic interpretation. It minimizes the Kullback-Leibler distance between the true distribution and the parametric family to which the approximation belongs, and therefore it is the quasi-maximum likelihood estimator in the spirit of White (1982).

Suppose one is interested in evaluating the effect of receiving or not receiving a binary treatment under the assumption that the treatment satisfies some form of exogeneity. In the counterfactual framework of program evaluation, an individual i has two potential outcomes: $Y_{i,1}$ in the presence of the treatment and $Y_{i,0}$ in the absence of the treatment. The individual treatment effect is then defined as $\Delta_i = Y_{i,1} - Y_{i,0}$. In a non-experimental setup, because an individual cannot be in both states, one does not observe $Y_{i,0}$ and $Y_{i,1}$ at the same time. Therefore, Δ_i is not observable and the problem of estimating individual treatment effect is essentially a missing data problem.

Although individual treatment effect is not identifiable, the average treatment effect (ATE) is identifiable. The literature on estimating the average treatment effect and other related concepts is abundant. For recent reviews on this topic, see Imbens (2004) and Heckman and Vytlacil (2005). Generally given certain assumptions, the average treatment effect, conditional on observed covariates X, is identified by

$$\text{ATE} = \text{E}\left(Y_1 - Y_0 | X\right) = \text{E}\left(\Delta | X\right),$$

where Y_1 and Y_0 are the distributions of potential outcomes with and without treatment respectively, and Δ is the distribution of individual treatment effect.

Although of fundamental importance, the ATE only reflects certain aspects of the treatment effect distribution. As Heckman et al. (1997) observed, some important questions cannot be answered by the ATE. For example, to use the median voter model, we need to know if more than 50% of the population will benefit from the program. Without knowing the distribution of individual treatment effect, we are unable to answer this question. Also, merely examining the ATE might sometimes lead to misleading conclusions. Suppose the treatment effect is heterogeneous such that the positive effects and negative effects offset each other and so that the ATE is roughly zero. Although a zero ATE does not imply zero treatment effects, sometimes researchers mistakenly draw this conclusion.

Since the individual treatment effect is generally not identifiable, some recent recent work has instead tried to estimate the distribution of potential outcomes. In an instrumental variables setting, Abadie et al. (2002) and Chernozhukov and Hansen (2005) investigate estimation of differences in quantiles of the potential outcome distributions. Under the unconfoundedness assumption, Firpo (2005) develops an estimator for quantile treatment effect. An alternative "Change-in-Change" method, recently proposed by Athey and Imbens (2006), estimates the entire counterfactual distribution nonparametrically. Both the quantile-based method and the Change-in-Change estimator focus on estimating the counterfactual outcome distribution rather than the distribution of individual treatment effect. Generally, the differences in the quantiles of two potential outcome distributions are different from the individual treatment effect as individual's rank within the two distributions may change due to the treatment. Only under the assumption of rank invariance is the quantile treatment effect equal to the individual treatment effect (see Firpo, 2005 and references therein).

In principle, the two marginal distributions of potential outcomes can be identified using one of the methods just discussed. However, even with complete knowledge of the two potential outcome distributions Y_0 and Y_1, we generally cannot infer Δ from them as we do not know their joint distribution (Y_0, Y_1). For example in a two-person economy, consider two

possible outcome distributions before and after a certain treatment. First, $Y_{1,0} = Y_{1,1} = 1$, $Y_{2,0} = Y_{2,1} = 2$, so that the individual treatment effect is $\Delta_1 = \Delta_2 = 0$. Second, $Y_{1,0} = 1, Y_{1,1} = 2$, and $Y_{2,0} = 2, Y_{2,1} = 1$, so that the treatment effect is $\Delta_1 = 1, \Delta_2 = -1$. In both cases, we have the same marginal distribution $Y_0 = [1, 2]$ and $Y_1 = [1, 2]$. Nonetheless, the individual treatment effects differ, depending on the correlation of individual outcomes, or their joint distribution, between two periods. In the absence of the rank invariance assumption, the knowledge of the marginal distributions is not sufficient to identify the distribution of the individual treatment effect.[1]

In this study, we present a method to approximate the distribution of individual treatment effects based on the standard ATE model. Consider the model

$$Y_i = f(X_i) + \Delta_i D_i + \varepsilon_i,$$

where $f(X_i)$ is a consistent estimator of the potential outcome in the absence of the treatment, D_i is the treatment indicator, and ε_i is an iid error which is independent of X_i. Denote $r_i = \Delta_i D_i + \varepsilon_i$. One can estimate the distribution of the error from the control group, and the distribution of $\Delta_i + \varepsilon_i$ from the treatment group. Under the unconfoundedness assumption that $(Y_0, Y_1) \perp D|X$, for the treatment group, r_i is the convolution of two independent random variables, Δ_i and ε_i. To estimate the target density Δ from r and ε is thus the classical deconvolution problem.

The deconvolution problem has been traditionally solved using the empirical characteristic function approach. For applications of this method in econometrics, see for example Horowitz and Markatou (1996) and Heckman et al. (1997). However, Carroll and Hall (1988) and Fan (1992) show that the convergence rate of optimal deconvolution can be slow, which makes it practically impossible to consistently estimate the deconvolution. Instead of trying to estimate the target distribution consistently, Carroll and Hall (2004) propose deconvolution methods to estimate a low-order approximation of the target density. Our

[1]Heckman et al. (1997) discuss how to obtain bounds of var(Δ) under certain statistical and behavioral assumptions.

proposed information-theoretic methods are in spirit close to their methods.

This paper makes two contributions. First, we propose an alternative deconvolution method, which our Monte Carlo simulations indicate compare favorably to existing methods, especially for the important case when the unknown target distribution deviates from the normal distribution. Only the first few moments of the convoluted data and the noise are required to implement the proposed method. Although it is close in spirit to the orthogonal series method in Carroll and Hall (2004), the proposed method offers some advantages: (i) the density estimate is positive everywhere, whereas the density estimates from the empirical characteristic function approach and Carroll and Hall's orthogonal series approximation sometimes produce negative densities; (ii) this method provides a well-defined quasi-maximum likelihood estimator that minimizes the Kullback-Leibler distance between the target density and the parametric family of the estimator (see White (1982) for a discussion of the quasi-maximum likelihood estimator and its asymptotic properties.)

Second, we use this new deconvolution estimator to approximate the entire distribution of the individual treatment effect. The distribution function has a simple functional form yet is flexible enough to impose few restrictions on the shape of the distribution. The proposed method extends the standard average treatment effect estimator in a straightforward manner by utilizing not only the ATE, but also implied higher-order moment information. Without structural assumptions on the distribution of the treatment effect, this method can estimate the treatment effect distribution even if it is entirely random or dependent on unknown covariates.

The next section presents the model and the method to estimate moments of the target distribution from the standard ATE model. Section 3 introduces the maximum entropy density, a method that constructs density estimate based on a given set of moment conditions. It then applies this method to approximation of deconvolution estimation. The asymptotic properties of the proposed methods are discussed. Section 4 uses Monte Carlo simulations to demonstrate the numerical performance of the proposed deconvolution method. It then applies this method to pseudo program evaluation data constructed from real data. Section

5 applies the proposed method to estimate the distribution of treatment effects of the U.S. Job Training Partnership Act. The final section draws conclusions.

2 The model

In this section, we show how to obtain consistent moment estimates of individual treatment effect distribution based on the standard ATE model. As is done in the ATE model, we assume that individual treatment effect is additive.

Assumption 2.1 The model is known to be

$$Y_i = f(X_i) + D_i\Delta_i + \varepsilon_i,\ i = 1, \ldots, n, \tag{1}$$

where $\{X_i\}$ is a sequence of uniformly bounded, fixed $1 \times m$ vectors, $f(X_i)$ is a consistent estimator of $Y_{i,0}|X_i$, and $\{\varepsilon_i\}$ is a sequence of iid random variables with $E(\varepsilon_i) = 0, E(\varepsilon_i^2) < \infty$.

Assumption 2.2 $X \perp \varepsilon$ and $\Delta \perp \varepsilon$.

Assumption 2.3 $(Y_0, \Delta) \perp D|X$.

Assumption 2.4 $0 < \Pr(D = 1|X) < 1$.

Assumption 2.1 ensures that $f(X_i)$ is a well-defined, consistent estimator of $Y_{i,0}$. Since $Y_1 = Y_0 + \Delta$, under Assumption 2.2, Assumption 2.3 is equivalent to the standard unconfoundedness assumption that $(Y_0, Y_1) \perp D|X$.[2] Assumption 2.4 is the overlap assumption that ensures the validity of extrapolation between the treatment and control groups. We do not impose any additional assumptions beyond those of the standard ATE model.[3] Although it is often (implicitly) assumed in the standard ATE model that the treatment effect is constant, we allow heterogeneous individual treatment effect in our model. If we are interested in estimating the average treatment effect, we can simplify the model to the standard ATE

[2]Sometimes a weaker mean independence assumption is assumed in the place of the unconfoundedness assumption. However, under the mean independence assumption, one cannot necessarily identify average treatment effects on all transformations for the original outcome.

[3]Imbens (2004) discusses the plausibility of the unconfoundedness and overlap assumptions.

model $Y_i = f(X_i) + D_i \mathrm{E}\Delta + \varepsilon_i$.

Following the standard practice, we assume that X is observed. However, we do not rule out the possibility that the individual treatment effect Δ is determined by a set of variables Z, provided that $Z \perp \varepsilon$. There is no restriction on the relationship between X and Z, so that all the following scenarios are possible: (1) X and Z are identical; (2) one is a strict subset of the other; (3) they are different and have a non-zero intersection; (4) they are mutually exclusive. When all variables in Z are observed, we can model $\Delta(Z_i)$ explicitly in principle. However, if some variables in Z are not observable, this method will suffer from omitted variable bias.

Instead, we propose a method that approximates the distribution of the individual treatment effect that does not require structural knowledge or assumptions on Δ. Our method is based on the higher order information of Δ and ε provided by the standard ATE model. Because the ATE model ignores information other than changes in the mean, Athey and Imbens (2006) propose a method where all changes in the distribution of Y_i across subpopulations are given a structural interpretation and used to estimate the entire counterfactual outcome distributions in the presence or absence of the treatment. Although their method identifies the outcome distributions, the distribution of individual treatment effect is not directly inferrable from the two marginal distributions, without knowledge about their joint distribution. Our approach is conceptually different in that we are trying to estimate the distribution of the individual effect $\Delta_i = Y_{i,1} - Y_{i,0}$ directly, rather than the two outcome distributions. Since consistent deconvolution estimation of Δ is practically insoluble, the goal of our method is to approximate Δ under a minimum set of assumptions and information requirements. Taking a middle ground between the standard ATE model that only uses the first moment of the outcome distributions, and the Athey and Imbens (2006)'s nonparametric model that uses all the changes between the subgroups, we base our analysis on the first few moments of $r_i = D_i\Delta_i + \varepsilon_i$ from the control and treatment group.

When $D_i = 0$ such that $r_i = \varepsilon_i$, we can estimate the moments of ε_i from the control group. On the other hand, when $D_i = 1$, $r_i = \Delta_i + \varepsilon_i$. Denote the k^{th} moment of Δ and ε

by μ_k and ν_k respectively. Since $\Delta \perp \varepsilon$, it follows that

$$\begin{aligned}
&\mathrm{E}\left(\Delta_i + \varepsilon_i\right)^k \\
=&\mathrm{E}\left\{\sum_{j=0}^{k} \frac{k!}{(k-j)!j!} \Delta_i^{k-j} \varepsilon_i^j\right\} \\
=&\sum_{j=0}^{k} \frac{k!}{(k-j)!j!} \mu_{k-j} \nu_j. \qquad (2)
\end{aligned}$$

Therefore, we can estimate the moments of Δ_i from the moments of ε_i and $\Delta_i + \varepsilon_i$. Denote $\hat{Y}_{i,0} = \hat{f}\left(X_i\right)$ from model (1), we obtain $\hat{r}_i = Y_i - \hat{Y}_{i,0}$. We further assume that

Assumption 2.5 $\mathrm{E}\left|\varepsilon_i\right|^{K+\delta} < \infty$ and $\mathrm{E}\left|\Delta_i\right|^{K+\delta} < \infty$, where K is a positive integer and $\delta > 0$.

The following Lemma establishes the consistency of the moment estimates.

Lemma 1. *Define, for* $k = 1, \ldots, K$,

$$\begin{aligned}
\hat{\nu}_k &= \frac{\sum_i^n \left(1 - D_i\right) \hat{r}_i^k}{\sum_i^n \left(1 - D_i\right)}, \\
\hat{\varpi}_k &= \frac{\sum_i^n D_i \hat{r}_i^k}{\sum_i^n D_i}.
\end{aligned}$$

Given Assumption 2.1 to 2.5,

$$\begin{aligned}
\left|\hat{\nu}_k - \nu_k\right| &\xrightarrow{as} 0, \\
\left|\hat{\omega}_k - \omega_k\right| &\xrightarrow{as} 0,
\end{aligned}$$

where $\omega_k = E\left(\Delta_i + \varepsilon_i\right)^k$.

All proofs are presented in Appendix.

When $k = 1$, if we substitute $\hat{\omega}_1$ and $\hat{\nu}_1$ into Equation (2) and solve for μ_1, we obtain $\tilde{\mu}_1 = \hat{\omega}_1 - \hat{\nu}_1 = \hat{\omega}_1$ since $\hat{\nu}_1 = 0$. When $k = 2$, substituting $\hat{\omega}_2$ and $\hat{\nu}_2$ into Equation (2), we obtain $\tilde{\mu}_2 = \hat{\omega}_2 - \hat{\nu}_2$. We can solve for higher order moments recursively in a similar manner.

The third and fourth moments are computed as

$$\tilde{\mu}_3 = \hat{\omega}_3 - 3\hat{\mu}_1\hat{\nu}_2 - \hat{\nu}_3,$$
$$\tilde{\mu}_4 = \hat{\omega}_4 - 6\hat{\mu}_2\hat{\nu}_2 - 4\hat{\mu}_1\hat{\nu}_3 - \hat{\nu}_4.$$

Note that all terms involving $\hat{\nu}_1$ disappear because, given that the estimates are consistent, $\mathrm{E}\hat{\nu}_1 = 0$.

The following theorem establishes the consistency of $\tilde{\mu}_k$, estimated moments of the treatment effect distribution, for $k = 1, \ldots, K$.

Theorem 2. *Given Assumption 2.1 to 2.6,* $|\tilde{\mu}_k - \mu_k| \xrightarrow{as} 0$ *for* $k = 1, \ldots, K$.

Certain restrictions on the estimated moments can be used as specification tests on the independence assumption $\Delta \perp \varepsilon$. For example, $\omega_2 = \mathrm{E}\left\{(\Delta + \varepsilon)^2\right\} = \mu_2 + \nu_2 + 2\mathrm{E}\left[\Delta\varepsilon\right]$ implies that $\mu_2 = \omega_2 - \nu_2 - 2\mathrm{E}\left[\Delta\varepsilon\right]$. Suppose Δ and ε are negatively correlated, we might have $\mu_2 = \omega_2 - \nu_2 - 2\mathrm{E}\left[\Delta\varepsilon\right] > 0$, but $\omega_2 - \nu_2 < 0$. If we incorrectly assume $\Delta \perp \varepsilon$, we would obtain $\mu_2 = \omega_2 - \nu_2 < 0$ so that the second moment of Δ is negative. Therefore, a negative $\tilde{\mu}_2$ clearly indicates a violation of Assumption 2.2. Similarly, if we find that $\tilde{\mu}_4 < 0$, we reject the independence assumption. In addition, the relationship between moments also provides testable restrictions. For example, for any standardized random variable, it is known that $\mu_4 > 1 + \mu_3^2$. Therefore, we can use this relationship between the third and fourth moments to test the "admissibility" of estimated moments.

We next proceed to construct an estimate of the target density Δ based on its first K estimated moments. Although there may exist an infinite number of distributions satisfying a given set of moment conditions, in the next section, we introduce an information-theoretic method of distribution construction, which leads to a unique, least biased distribution estimator based some moment conditions.

3 Information-Theoretic Deconvolution

In this section, we present a deconvolution estimator that is a low-order approximation to the target density. This density estimator is obtained by maximizing entropy subject to a given set of moment conditions. With a slight abuse of notation in this section, we denote the target density to estimate by $p_0(X)$. For the maximum entropy density estimation in section 3.1, we assume that the first K moments of X are known or can be consistently estimated. For the deconvolution problem in section 3.2, we do not observe X, but observe data on Y and $Z = X + Y$, where $X \perp Y$.

3.1 Maximum Entropy Density

Suppose one is to construct a probability distribution using limited information, in our case, moments. The principle of maximum entropy (maxent) states that one should choose the probability distribution, consistent with given information, that maximizes Shannon's measure of entropy. According to Jaynes (1957), the resulting maximum entropy distribution is "uniquely determined as the one which is maximally noncommittal with regard to missing information, and that it agrees with what is known, but expresses maximum uncertainty with respect to all other matters."

The maxent density is obtained by maximizing Shannon's information entropy measure

$$W = -\int p(x) \log p(x)\, dx,$$

subject to K known moment conditions

$$\int x^k p(x) dx = \mu_k,\ k = 1, \ldots K.$$

Using Lagrange's method, we can obtain a unique global maximum entropy solution to this

problem:

$$p(x,\theta) = \frac{\exp\left(-\sum_{k=1}^{K}\theta_k x^k\right)}{\int \exp\left(-\sum_{k=1}^{K}\theta_k v^k\right) dv}, \tag{3}$$

where θ_k is the Lagrange multiplier for the k^{th} moment constraint and $\theta = (\theta_1, \ldots, \theta_K)$.[4]

Zellner and Highfield (1988) and Wu (2003) discuss the estimation of maxent density subject to moment constraints. Generally this problem has no analytical solution. To solve for the Lagrange multipliers, we use Newton's method to iteratively update

$$\theta^{(1)} = \theta^{(0)} + \mathcal{H}^{-1} b,$$

where $b_k = \mu_k - \int x^k p(x,\theta)\, dx$, and $\mathcal{H}$ is the K by K Hessian matrix of the form, for $1 \leq k, j \leq K$,

$$\mathcal{H}_{kj} = \int x^{k+j} p(x,\theta)\, dx - \int x^k p(x,\theta)\, dx \int x^j p(x,\theta)\, dx. \tag{4}$$

This maximum entropy method is equivalent to a maximum likelihood approach where the likelihood function is defined over the exponential distribution and therefore the estimated coefficients are asymptotically consistent and efficient (see Golan et al., 1996 for a discussion of the duality between these two approaches). Most of the well-known distributions used in mathematical statics may be characterized as maxent densities subject to certain moment constraints. For example, the normal distribution is a maxent density with characterizing moments x and x^2, the gamma distribution is characterized by x and $\log x$ for $x > 0$, and the beta distribution by $\log(x)$ and $\log(1-x)$ for $0<x<1$.

Next we derive the asymptotic properties of $p(x,\theta)$ defined in Equation (3). Suppose $X_i, i = 1, \ldots, n$, is an iid sample from a distribution with density function $p_0(x)$. As discussed above, the maxent density is equivalent to the MLE of $p(x,\theta)$. When the partial derivatives

[4]The existence and uniqueness of the solution is guaranteed, because the Hessian of the optimization problem, which is given by Equation (4), is positive definite.

of the log-likelihood function exist, we define the matrices

$$A_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log p(X_i, \theta)}{\partial\theta_k \partial\theta_j},$$
$$B_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log p(X_i, \theta)}{\partial\theta_k} \frac{\partial \log p(X_i, \theta)}{\partial\theta_j}.$$

Assuming their expectations exist, we define

$$A(\theta) = \mathrm{E}[A_n(\theta)],$$
$$B(\theta) = \mathrm{E}[B_n(\theta)].$$

The following theorem provides the asymptotic properties of the maxent density estimate.

Assumption 3.1 The independent random sample $X_i, i = 1, \ldots, n$, has a common joint distribution function P_0 on Ω, a measurable Euclidean space, with measurable density $p_0 = dP_0/dv$, which is continuously differentiable, positive and uniformly bounded on its support.

Assumption 3.2 The family of distributions $P(x, \theta)$ has density functions $p(x, \theta) = dP(x, \theta)/dv$ which are measurable in x for every θ in Θ, a compact subset of a K-dimensional Euclidean space, and continuous in θ for every x in Ω.

Assumption 3.3 (a) $\mathrm{E}[\log p_0(X_i)]$ exists and $|\log p(x, \theta)| \leq m(x)$ for all θ in Θ, where m is integrable with respect to P_0; (b) $I(p_0 : p, \theta)$ has a unique minimum at θ^*, which is interior to Θ.

Assumption 3.4 $\partial \log p(x, \theta)/\partial\theta_k$, $k = 1, \ldots, K$, are measurable functions of x for each θ in Θ and continuously differentiable function of θ for each x in Ω.

Assumption 3.5 (a) $|\partial^2 \log p(x, \theta)/\partial\theta_j \partial\theta_k|$ and $|\partial \log p(x, \theta)/\partial\theta_j \cdot \partial \log p(x, \theta)/\partial\theta_k|$, $j, k = 1, \ldots, K$ are dominated by functions integrable with respect to P_0 for all x in Ω and θ in Θ; (b) $B(\theta^*)$ is non-singular; (c) θ^* is a regular point of $A(\theta)$.

Note that due to the duality between the maxent method and the MLE approach, these assumptions are essentially identical to the regularity conditions for the MLE. Denote $\hat{\theta}$ the

maximum likelihood estimates, which are equivalent to the maxent estimates subject to the first K moment constraints.

Theorem 3. *Given the regularity conditions 3.1 to 3.5,*

$$\sqrt{n}\left(\hat{\theta}-\theta^{*}\right)\sim\mathcal{N}\left(0,A\left(\theta^{*}\right)^{-1}B\left(\theta^{*}\right)A\left(\theta^{*}\right)^{-1}\right).$$

Moreover, $A_n\left(\hat{\theta}\right)\overset{as}{\to}A\left(\theta^{*}\right)$ *and* $B_n\left(\hat{\theta}\right)\overset{as}{\to}B\left(\theta^{*}\right)$ *element by element.*

The following collorary establishes the asymptotic distribution of the estimated density.

Corollary 4. *Given the regularity conditions 3.1 to 3.5,*

$$\sqrt{n}\left(p\left(x,\hat{\theta}\right)-p\left(x,\theta^{*}\right)\right)\sim\mathcal{N}\left(0,\mathcal{G}\left(x,\theta^{*}\right)A\left(\theta^{*}\right)^{-1}B\left(\theta^{*}\right)A\left(\theta^{*}\right)^{-1}\mathcal{G}\left(x,\theta^{*}\right)'\right),$$

where $\mathcal{G}\left(x,\theta^{*}\right)=\nabla_{\theta}p\left(x,\theta^{*}\right)$ *with* $\mathcal{G}_k\left(x,\theta^{*}\right)=\left\{x^k-\mu_k\left(\theta^{*}\right)\right\}p\left(x,\theta^{*}\right),\ k=1,\ldots,K$. *Moreover,* $\mathcal{G}_k\left(x,\hat{\theta}\right)\overset{as}{\to}\mathcal{G}_k\left(x,\theta^{*}\right)$.

Since θ^{*} is not observable, we replace it with its estimates in the calculation of covariance matrix. Denote $\hat{\mu}_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$ and $\mu_k\left(\theta\right)=\int x^k p\left(x,\theta\right)dx$. Note that

$$\log p\left(X_i,\theta\right)=-\sum_{k=1}^{K}\theta_k X_i^k-\log\int\exp\left(-\sum_{k=1}^{K}\theta_k v^k\right)dv.$$

It follows that

$$\begin{aligned}\frac{\partial\log p\left(X_i,\theta\right)}{\partial\theta_k}&=-X_i^k+\int\frac{v^k\exp\left(-\sum_{k=1}^{K}\theta_k v^k\right)}{\int\exp\left(-\sum_{k=1}^{K}\theta_k v^k\right)dv}dv\\&=-X_i^k+\mu_k\left(\theta\right),\end{aligned}$$

and

$$\frac{\partial^2 \log p\left(X_i,\theta\right)}{\partial\theta_k\partial\theta_j} = \int v^k \left\{ -v^j p\left(v,\theta\right) + \frac{\exp\left(-\sum_{k=1}^K \theta_k v^k\right) \int v^j \exp\left(-\sum_{k=1}^K \theta_k v^k\right) dv}{\left(\int \exp\left(-\sum_{k=1}^K \theta_k v^k\right) dv\right)^2} \right\} dv$$
$$= -\mu_{k+j}\left(\theta\right) + \mu_k\left(\theta\right)\mu_j\left(\theta\right).$$

We then have

$$\begin{aligned} A_n &= -\mu_{k+j}\left(\hat{\theta}\right) + \mu_k\left(\hat{\theta}\right)\mu_j\left(\hat{\theta}\right) \\ &= -\mu_{k+j}\left(\hat{\theta}\right) + \hat{\mu}_k\hat{\mu}_j, \\ B_n &= \frac{1}{n}\sum_{i=1}^n \left\{X_i^k - \mu_k\left(\hat{\theta}\right)\right\}\left\{X_i^j - \mu_j\left(\hat{\theta}\right)\right\} \\ &= \hat{\mu}_{k+j} - \hat{\mu}_k\mu_j\left(\hat{\theta}\right) - \hat{\mu}_j\mu_k\left(\hat{\theta}\right) + \mu_k\left(\hat{\theta}\right)\mu_j\left(\hat{\theta}\right) \\ &= \hat{\mu}_{k+j} - \hat{\mu}_k\hat{\mu}_j. \end{aligned}$$

The last equality hold since $\mu_k\left(\hat{\theta}\right) = \hat{\mu}_k$ for $1 \leq k \leq K$.

Since the density (3) only approximates the true distribution $p_0\left(x\right)$, generally $\mu_k\left(\hat{\theta}\right) \neq \hat{\mu}_k$, except for $1 \leq k \leq K$ because the maxent density matches its first K moments to those of $p_0\left(x\right)$. Therefore generally $\left|A_n\right| \neq \left|B_n\right|$, as $\mu_{k+j}\left(\hat{\theta}\right) \neq \hat{\mu}_{k+j}$ for $k+j > K$. This discrepancy between the information matrix and the outer product of the gradient is discussed in details in White (1982). Define the Kullback-Leibler distance between two densities $f\left(x\right)$ and $g\left(x\right)$

$$I\left(f:g\right) = \int f\left(x\right)\log\left\{f\left(x\right)/g\left(x\right)\right\}dx,$$

which is non-negative and equal to zero if and only if $f\left(x\right) = g\left(x\right)$ everywhere. The estimated maxent density is well-defined and it is in fact the unique quasi-maximum likelihood estimator that minimizes the Kullback-Leibler distance between the target density and the parametric family to which the maxent density belongs.

Theorem 5. *Define* $p(x,\theta) = \exp\left(-\sum_{k=1}^{K}\theta_k x^k\right) / \int \exp\left(-\sum_{k=1}^{K}\theta_k v^k\right) dv$. *The maxent estimate* $p\left(x,\hat{\theta}\right)$ *obtained by maximizing the entropy subject to the first* K *moments of* $p_0(x)$ *minimizes the Kullback-Leibler distance* $I(p_0 : p(x,\theta)) = \int p_0(x) \log\{p_0(x)/p(x,\theta)\}\, dx$.

Theoretically, we can approximate a distribution arbitrarily well by increasing the number of moments (Cobb et al., 1983 and Barron and Sheu, 1991). However in practice, empirical moments higher than the fourth order are rarely used or needed. Because higher-order sample moments are sensitive to outliers, so are density estimators involving higher-order moments. In this study, we use the first four moments for density approximation. The maxent density then takes the form

$$p(x,\theta) = \frac{\exp\left(-\sum_{k=1}^{4}\theta_k x^k\right)}{\int \exp\left(-\sum_{k=1}^{4}\theta_k v^k\right) dv}. \tag{5}$$

Although simple in functional form, this maxent density based on the first four moments is flexible enough to approximate distributions of various shapes. For example, it allows for both uni-modal and multi-modal distributions.[5]

3.2 Deconvolution Approximation

In this section, we present the maxent approximation to the target density $p_0(x)$ through the deconvolution of $Z_i = X_i + Y_i, i = 1,\ldots,n$. Suppose we also observe a separate sample $Y_t, t = 1,\ldots,T$. In addition to Assumption 3.1 to 3.5 on X_i, we also assume

Assumption 3.6 The independent random sample $Y_i, i = 1,\ldots,n$, has a common joint distribution function $P(y)$ on Ψ, a measurable Euclidean space; $E|Y_i|^{2K+\delta} < \infty$, $\delta > 0$.

Assumption 3.7 $X \perp Y$.

Since X is not observed, we use moment information from Y and Z to approximate the distribution of X. Define $\hat{\mu}_k(z) = \frac{1}{n}\sum_{i=1}^{n} Z_i^k$ and $\hat{\mu}_k(y) = \frac{1}{T}\sum_{t=1}^{T} Y_t^k$. Given $\hat{\mu}_k(y)$ and $\hat{\mu}_k(z)$, we can estimate $\tilde{\mu}_k(x)$, the moments for the target density $p_0(x)$, using the recursive method described in Section 2. We then use the maxent density method described in the previous section to approximate the target density based on $\tilde{\mu}(x) = \{\tilde{\mu}_1(x),\ldots,\tilde{\mu}_K(x)\}$. De-

[5]Cobb et al. (1983) discuss the relationship between number of modes and the moment conditions.

note the resulting maxent density by $p\left(x,\tilde{\theta}\right)$. The following theorem establishes its asymptotic normality.

Theorem 6. *Given assumption 3.1 to 3.7,*

$$\sqrt{n}\left(\tilde{\theta}-\theta^{*}\right)\sim\mathcal{N}\left(0,A\left(\theta^{*}\right)^{-1}B\left(\theta^{*}\right)A\left(\theta^{*}\right)^{-1}\right).$$

Moreover, $A_n\left(\tilde{\theta}\right)\overset{as}{\to}A\left(\theta^{*}\right)$ *and* $B_n\left(\tilde{\theta}\right)\overset{as}{\to}B\left(\theta^{*}\right)$ *element by element.*

The asymptotic normality of the estimated maxent density follows immediately by the following collorary.

Corollary 7. *Given assumption 3.1 to 3.7,*

$$\sqrt{n}\left(p\left(x,\tilde{\theta}\right)-p\left(x,\theta^{*}\right)\right)\sim\mathcal{N}\left(0,\mathcal{G}\left(x,\theta^{*}\right)A\left(\theta^{*}\right)^{-1}B\left(\theta^{*}\right)A\left(\theta^{*}\right)^{-1}\mathcal{G}\left(x,\theta^{*}\right)'\right),$$

where $\mathcal{G}\left(x,\theta^{*}\right)=\nabla_{\theta}p\left(x,\theta^{*}\right)$ *with* $\mathcal{G}_k\left(x,\theta^{*}\right)=\left\{x^k-\mu_k\left(\theta^{*}\right)\right\}p\left(x,\theta^{*}\right)$, $k=1,\ldots,K$. *Moreover,* $\mathcal{G}_k\left(x,\tilde{\theta}\right)\overset{as}{\to}\mathcal{G}_k\left(x,\theta^{*}\right)$.

Although $A\left(\theta^{*}\right)$ and $B\left(\theta^{*}\right)$ are not observable, they can be estimated consistently from the data by

$$\begin{aligned}A_n&=-\mu_{k+j}\left(\tilde{\theta}\right)+\tilde{\mu}_k\tilde{\mu}_j,\\B_n&=\tilde{\mu}_{k+j}-\tilde{\mu}_k\tilde{\mu}_j,\end{aligned}$$

for $1\leq j,k\leq K$.

The proposed method offers several advantages over conventional methods. The conventional deconvolution method is based on the estimation of the empirical characteristic function of X_i, which is computed as the ratio of the estimated empirical characteristic functions of Z_i and Y_i. Fourier inversion is then employed to convert the estimated characteristic function to a density. This method has two difficulties. First, the nonparametric estimation of empirical characteristic function may be sensitive to the selection of bandwidth. Second,

the density obtained from the empirical characteristic function through Fourier inversion is not guaranteed to be positive. Carroll and Hall (2004) propose two alternative methods, involving kernel and orthogonal series methods, to obtain a low-order approximation of the target density rather than its consistent estimate. Their methods only require knowledge of the first few moments of Y and Z. Although our estimator is also a low-order approximation to the target density, it is equivalent to the quasi-maximum likelihood estimator and minimizes the Kullback-Leibler distance between the target density and the parametric family (3) to which the proposed estimator belongs.

Lastly, we note that for the deconvolution approximation of the treatment effect distribution discussed in Section 2, a complication arises as we do not directly observe $r_i = D_i\Delta_i + \varepsilon_i$, though we have an estimate from a regression. Nonetheless, according to White and McDonald (1980), Corollary 2 ensures that under regularity conditions 2.1 to 2.5, when we replace r_i by $\hat{r}_i$, the estimated densities follow the same asymptotic distribution as prescribed by Corollary 6.

We use Equation (5) for the deconvolution approximation in this study. Because the maxent density subject to the first two moments is the normal density, if we use only the first two moments, the proposed estimator is equivalent to a random coefficient model where the the random coefficients are distributed normally. We also note that the Pearson family distributions can be completely characterized by their first four moments. Biddel et al. (2003) use the Pearson family to approximate the treatment distribution. However, their estimator only allows uni-modal distributions, which limits the applicability of their method. Our proposed maxent estimator is more flexible and imposes few restrictions on the shape of the distribution.

4 Numerical Performance

In this section, we first present Monte Carlo simulations on the numerical performance of the proposed deconvolution approximation. We then apply the estimator to a pseudo program evaluation data to assess its performance of approximating the distribution of individual

treatment effect.

4.1 Monte Carlo Simulations

Following Carroll and Hall (2004), we set $\mathrm{E}\Delta = \mathrm{E}\varepsilon = 0$, $\sigma_\Delta^2 = 4/3$ and $\sigma_\varepsilon^2 = 1/3$ for all the distributions. For the target density Δ, we consider the normal distribution, the skew-normal distribution with skewness parameter 5 and density function $2\,\phi(x)\,\Phi(5x)$,[6] and a normal mixture with equal probability of $\mathcal{N}(1.2, 0.5)$ and $\mathcal{N}(-1.2, 0.5)$. The normal distribution is the usual baseline case; the skew-normal distribution is used to show the performance of the estimator under skewness; the mixed normal case demonstrates that the estimator works for bi-modal distribution. The distributions of the error term include the normal distribution and the uniform distribution in $[-1, 1]$. All distributions are scaled to have the prescribed variances.

In each experiment, we generate two random vectors Δ and ε of size n, and then set $r_i = \Delta_i + \varepsilon_i$, $i = 1, \ldots, n$. The first four moments of r_i and ε_i are used in the deconvolution to approximate the distribution of the target density Δ. We experiment with two sample sizes, $n = 250$ and $n = 500$. Each experiment is repeated 500 times. For comparison, we also report the results of Carroll and Hall's (2004) orthogonal series estimator, which they show outperforms traditional methods. Their method approximates the target distribution using the Gram-Charlier expansion through Hermite polynomials, which is also based on the estimated moments discussed earlier. One drawback of the Gram-Charlier expansion, like the closely-related Edgeworth expansion, is that it does not guarantee positive densities.[7]

We use the integrated squared errors $\int \{p_0(x) - \hat{p}(x)\}^2 dx$ to gauge the goodness-of-fit, where $p_0(x)$ and $\hat{p}(x)$ are the theoretical and estimated target densities respectively. The results are reported in Table 1. Also reported is the mean squared errors (MSE) of the inter-quartile range, an alternative robust measure of dispersion.

For $n = 250$, the maxent estimator outperforms the orthogonal series except when the

[6] With a skewness parameter α, the skew-normal distribution with density $2\phi(x)\,\Phi(\alpha x)$ has mean $\mu = \alpha\sqrt{2\pi}/\sqrt{(1+\alpha^2)}$ and variance $\sigma^2 = 1 - \mu^2$.

[7] In our simulations, the orthogonal series estimates are set to be zero whenever negative densities occur.

target density is normal. The better performance of the orthogonal series estimator for the normal density is to be expected as the Gram-Charlier expansion works best when the underlying distribution is normal or near-normal. However in empirical work, we have no reason to assume an unknown distribution is normal. As the target density deviates from the normal distribution, the performance of the orthogonal series estimator deteriorates rapidly. When the target density is skewed as in the case of skewed-normal distribution, the integrated squared errors of the maxent estimator are only about 85% of those of orthogonal series estimator. For the bi-modal mixed-normal density, this ratio is below 25%. The performance of both deconvolution estimators is slightly better when the noise distribution is uniform rather than normal, reflecting the difficulty in filtering out Gaussian noise.

In columns (b) of Table 1, we report the mean squared errors (MSE) of the inter-quartile range of the two estimated densities. For the normal distribution, the orthogonal series estimator performs better as it is implicitly based on normal distribution. For the skew-normal distribution, the MSE of the maxent estimator is about 75% of that of the orthogonal series estimator. For the bi-modal mixed normal distribution, the MSE of the maxent estimator is less than 10% of that of the orthogonal series estimator.

The results for $n = 500$ are qualitatively similar to those for $n = 250$. Now the two estimators perform nearly equally well even when the target density is the normal distribution, where the orthogonal series estimator is expected to be competitive.

Overall, our experiments show that the proposed estimator is able to approximate the target density well and often substantially outperforms existing methods, especially for the important cases when the target density deviates from the normal distribution.

4.2 Experiments with Pseudo Program Evaluation Data

In this section, we apply the proposed method to examine a pseudo program evaluation problem constructed from real data. The data are extracted from the Current Population Survey Outgoing Rotation Group file of April, 2004. Our sample is restricted to prime-aged, full-time workers with hourly wage between 5 and 100 dollars. We use the logarithm of wages

in the experiments.

We construct our pseudo program data according to the standard assumptions of the popular difference-in-difference estimator, where there are two periods, before and after the treatment, and two randomly assigned groups, the control and treatment group. Hence, to each individual in the sample, we randomly assign two independent Bernoulli variables with a 50% probability of success, one indicating "Time" (T_i) and the other indicating "Group" (G_i). Thus, we randomly divide the sample into four mutually exclusive and exhaustive groups: a control group in the first period ($T_i = 0$, $G_i = 0$), a control group in the second period ($T_i = 1$, $G_i = 0$), a treatment group in the first period ($T_i = 0$, $G_i = 1$), and a treatment group in the second period ($T_i = 1$, $G_i = 1$). The sample has 4,266 observations, so that each group averages slightly over 1,000 observations. The treatment indicator is $D_i = T_i G_i$, with $D_i = 1$ if the individual receives the treatment.

To introduce a time effect and a group effect to our data, we add 0.1 to the wage of each observation with $T_i = 1$ and 0.1 to that of those with $G_i = 1$. We then regress the constructed wage on a vector of social-economic control variables, including: age, age square, education, education square, sex, union status, plus the time and group dummies. A normally distributed error term with mean zero and standard deviation 0.1 is added to the fitted wage from the regression. This constructed wage has a mean 2.59 and a standard deviation 0.27. Lastly, for each treated individual (those in the treatment group of the second period), we add a hypothetical treatment effect to the constructed wage. We experiment with different signal/noise ratios in four experiment designs.

To summarize, the experiment procedure is

1. For individual i in the sample, assign random group and time status G_i and T_i;

2. Set $w_i = w_{i,0} + 0.1G_i + 0.1T_i$, where $w_{i,0}$ is logarithm of original wage rate of the i^{th} individual in the sample;

3. Regress w_i on the social-economic control variables and time and group dummies using the OLS, denote the fitted value $\hat{w}_i$;

4. Set $w_i^* = \hat{w}_i + D_i\Delta_i + \varepsilon_i$, where $D_i = T_iG_i$, Δ_i is the hypothetical individual treatment effect, and ε_i is a normal random error with mean zero and variance 0.1.

We use the standard ATE model as the first step of our estimation:

$$w_i^* = \beta_0 + X_i\beta + \alpha_1 G_i + \alpha_2 T_i + D_i\bar{\Delta} + \varepsilon_i, \tag{6}$$

where X_i is a vector of social-economic characteristics, α_1 and α_2 capture the group and time effect, and $\bar{\Delta}$ is the average treatment effect. By construction, the treatment status D_i is independent of the treatment effect Δ_i, and the error term is independent of the treatment status and individual treatment effect. After we compute the moments of the treatment effect distribution from the residuals of model (6). We estimate the treatment effect distribution using the maxent density based on the estimated moments.

In our experiments, we consider four different designs for the distribution of individual treatment effect. First, we randomly generate treatment effects distributed according to $\mathcal{N}(0.1, 0.1)$. The top panel of Figure 1 shows the histogram of the randomly generated treatment effects used in this experiment and the estimated maxent density. The estimated density tracks the data closely.

The second experiment involves a non-normal random treatment effect distribution. The data are generated according to the log-normal distribution, the exponential of which has mean -2 and standard deviation 0.5. The generated random effect sample for the treatment group has a mean 0.15 and standard deviation 0.08. The results are reported in the bottom panel of Figure 2. The approximation is surprisingly good even though the maxent density subject to the first four moments does not nest the log-normal distribution.[8]

In the third experiment, we assume that individual treatment effect depends on years of schooling. We generate a non-random heterogeneous treatment effect distribution according

[8]The log-normal distribution is a maxent density characterized by moments of $\log(x)$ and $\log(x)^2$ for $x > 0$.

to the following hypothetical formula

$$\Delta_i = 0.1 + 0.1 \times educ_i - 0.01 \times educ_i^2,$$

where *educ* is the years of schooling. The generated treatment sample has mean 0.17 and standard deviation 0.17. The top panel of Figure 3 plots the estimated treatment effect distribution, which is very close to the data used in the experiment. The two modes correspond to the clusters of high school and college graduates.

In the last experiment, we assume that individual treatment effect depends on the size of the local labor market, which is captured by the population size of Metropolitan Statistical Area (MSA). In addition, we allow random noise in the heterogeneous treatment effect. The key difference from the previous example is that this treatment effect depends on an unknown variable and therefore cannot be modeled directly even if we know the true data generating process. We generate this noisy heterogeneous treatment effect using the formula

$$\Delta_i = 0.05 \times \log(MSA_i) + v_i,$$

where MSA_i is the size of the metropolitan statistical area in which individual i lives and v_i is an iid random error generated according to a normal distribution with mean zero and variance 0.1. This constructed individual treatment effect has a mean 0.27 and standard deviation 0.06. The estimated distribution of the treatment effect is plotted in the bottom panel of Figure 2. The estimation captures the underling distribution remarkably good, even though the signal/noise ratio is substantially smaller in this experiment than in the previous one.

The above experiments show that the proposed method is able to approximate the distribution of individual treatment effect of various types with a simple functional form. An important feature of this method is that one can approximate the distribution closely based on the higher-order information implied by the ATE model even if the heterogeneous treatment effect depends on variables not included in the first step ATE model and unknown to

the researchers.

5 Empirical Application

In this section, we use the proposed estimator to estimate the impact distribution on earnings by the U.S. Job Training Partnership Act (JTPA), a large scale U.S. training program. We use the 4,317 observations in the recent National JTPA Study (NJS). Heckman et al. (1997) use the same dataset to estimate heterogeneous treatment effects. In the JTPA program, approved applicants were randomly assigned to treatment or control group. The treatment included classroom training, on-the-job training and job search assistance to the disadvantage. The control group was prohibited from receiving JTPA services for 18 months. Following Heckman et al. (1997), we focus on the earnings of adult women, where the outcome variable is individual earnings 18 months after the treatment.

We consider three specifications:

1. Unconditional: we compute the unconditional moments of the treatment effect.

2. Conditional: we estimate the moments of treatment effect distribution conditional on some observable social-economic control variables, including: age, sex, education, race, marital status, number of children, total family income, past work and welfare history, and experiment site dummies.

3. Conditional with Interactions: in addition to the variables in (2), we include interactions between the treatment status and age and education.[9]

The estimated moments of the treatment distribution are reported in Table 2. Consistent with Heckman et al. (1997), all the even moments are positive and therefore we do not reject the independence assumption $\varepsilon \perp \Delta$ based on estimated moments. Also, the estimated standardized moments satisfy the relationship $\mu_4 > 1 + \mu_3^2$, a necessary condition for

[9]The regression results are available from the authors upon request.

all distributions. All three estimated variances are statistically different from zero, indicating that the treatment effect is heterogeneous. The conditional variance is slightly smaller than the unconditional one. Adding the interaction terms with the treatment status barely changes the variance. Statistical tests do not reject the hypothesis that the three variances are identical, indicating that most of the variation in the treatment effect distribution cannot be explained by the covariates used in the interaction terms of our model. Therefore in this case, the observed individual characteristics offers little insight into the distribution of individual treatment effect. In contrast, the proposed estimator is suitable for this task as it imposes few restrictions on the shape of the distribution and is able to approximate the underlying distribution well, whether the distribution is entirely random or dependent on unknown factors.

Since all three specifications produce similar results, we focus our discussion on conditional moment estimates without interactions. Figure 3 plots the estimated treatment effect distribution (solid) with the asymptotic standard errors (dotted). The shape of the distribution is close to the one reported in Heckman et al. (1997), which use the empirical characteristic function approach for deconvolution. Also plotted is a normal distribution with the same mean and variance. The treatment distribution is more concentrated than the normal but skewed to the right. The Jarque and Bera test of normality rejects the hypothesis of normality at the 1% level. Hence, the conventional error component estimator assuming a normal treatment effect distribution will fail to capture some important features of the underlying impact distribution.

We can use this estimated treatment distribution to answer some interesting questions about the distributional impact of the program. For example, the median impact is \$327, less than the average treatment effect of \$844, suggesting that a relatively small group of people might have received disproportionately large benefits from this program. According to the estimated treatment effect distribution, 51% of the sample benefit from this program.[10]

[10]Using unconditional estimates, Heckman et al. (1997) report that 56% of the population benefit from this program.

6 Conclusion

The commonly used estimators for program evaluation focus on the average treatment effect. However, these estimators fail to capture important features of the distribution of heterogenous treatment effects, knowledge of which is instrumental in answering important policy questions.

In this study, we propose an information-theoretic deconvolution method to approximate the entire distribution of treatment effect. The method uses higher-order moment information implied by the standard average treatment effect estimator and estimates a flexible distribution using a maximum entropy method. By so doing, one can estimate the distribution of treatment effect, even if it is entirely random or dependent on unknown covariates. Monte Carlo and numerical examples demonstrate the effectiveness of the proposed estimator as a general deconvolution method and its promising performance when applied to a program evaluation problem.

We apply the proposed method to data from the JTPA experimental training program to estimate the distribution of the treatment effects on individual earnings. Our results suggest that little variation in the individual treatment can be explained by observables, highlighting the importance of modeling the treatment effect distribution as a flexible random process. Slightly more than 50% of the treated population is projected to benefit from this training program. We find that the impact distribution is substantially right-skewed, so that the average treatment effect is 2.6 times larger than the median treatment effect. Hence, examining the entire distribution provides valuable information that is not captured by the average effect.

Appendix

Proof of Theorem 1. Under Assumption 2.1, 2.5 and 2.6, we have for integer k with $1 \leq k \leq K$,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^k \xrightarrow{as} \mathrm{E}\varepsilon_i^k,$$
$$\frac{1}{n}\sum_{i=1}^{n}(1-D_i) \xrightarrow{as} \Pr(D_i=0) > 0.$$

Then under Assumption 2.3, it follows that

$$\frac{\sum_{i=1}^{n}(1-D_i)\,\hat{r}_i^k}{\sum_{i=1}^{n}(1-D_i)} \xrightarrow{as} \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^k \Pr(D_i=0)}{\Pr(D_i=0)} \xrightarrow{as} \mathrm{E}\varepsilon_i^k.$$

Similarly, under Assumption 2.1 to 2.6, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\Delta}_i+\hat{\varepsilon}_i\right)^k \xrightarrow{as} \mathrm{E}\left(\Delta_i+\varepsilon_i\right)^k,$$
$$\frac{1}{n}\sum_{i=1}^{n}D_i \xrightarrow{as} \Pr(D_i=1) > 0.$$

It follows that

$$\frac{\sum_{i=1}^{n}D_i\hat{r}_i^k}{\sum_{i=1}^{n}D_i} \xrightarrow{as} \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\Delta}_i+\hat{\varepsilon}_i\right)^k \Pr(D_{i=1})}{\Pr(D_{i=1})} \xrightarrow{as} \mathrm{E}\left(\Delta_i+\varepsilon_i\right)^k. \qquad \square$$

Proof of Corollary 2. Substituting $\hat{\omega}_k$ and $\hat{\nu}_k$ into the left and right hand side of Equation (2) and applying Theorem 1 gives the results immediately. $\square$

Proof of Theorem 3. The detail proof can be found in White (1982). Basically, given assumption 3.1 to 3.5,

$$\mathrm{E}\nabla_\theta \log p(X_i,\theta^*) = 0.$$

Applying Taylor's expansion to $\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta \log p\left(X_i,\hat{\theta}\right)=0$ yields

$$\sqrt{n}\left(\hat{\theta}-\theta^*\right)+A\left(\theta^*\right)^{-1}n^{-1/2}\sum_{i=1}^{n}\nabla_\theta \log p\left(X_i,\hat{\theta}\right)\xrightarrow{P}0.$$

It then follows that

$$\sqrt{n}\left(\hat{\theta}-\theta^*\right)\xrightarrow{d}\mathcal{N}\left(0,A\left(\theta^*\right)^{-1}B\left(\theta^*\right)A\left(\theta^*\right)^{-1}\right).$$

Given Assumption 3.5, $A_n\left(\theta^*\right)\xrightarrow{as}A\left(\theta^*\right)$. Since $\hat{\theta}\xrightarrow{P}\theta^*$ as is proved above, we then have

$$A_n\left(\hat{\theta}\right)\xrightarrow{as}A\left(\theta^*\right).$$

Similarly, we have $B_n\left(\hat{\theta}\right)\xrightarrow{as}B\left(\theta^*\right)$. □

Proof of Corollary 4. This corollary is a direct application the delta method. □

Proof of Theorem 5. This theorem is implied by Csiszár's (1975) information projection theorem. Since the densities are positive, we have

$$\log\frac{p_0\left(x\right)}{p\left(x,\theta\right)}=\log\frac{p_0\left(x\right)}{p\left(x,\hat{\theta}\right)}+\log\frac{p\left(x,\hat{\theta}\right)}{p\left(x,\theta\right)}.$$

Taking expectation with respect to $p_0(x)$ on both sides of yields

$$\int p_0\left(x\right)\log\frac{p_0\left(x\right)}{p\left(x,\theta\right)}dx=\int p_0\left(x\right)\log\frac{p_0\left(x\right)}{p\left(x,\hat{\theta}\right)}dx+\int p_0\left(x\right)\log\frac{p\left(x,\hat{\theta}\right)}{p\left(x,\theta\right)}dx,$$

which is equivalent to

$$D\left(p_0\left(x\right):p\left(x,\theta\right)\right)=D\left(p_0\left(x\right):p\left(x,\hat{\theta}\right)\right)+\int p_0\left(x\right)\log\frac{p\left(x,\hat{\theta}\right)}{p\left(x,\theta\right)}dx.$$

Note that

$$
\begin{aligned}
&\int p_0(x) \log \frac{p\left(x, \hat{\theta}\right)}{p(x, \theta)} dx \\
=&\int p_0(x) \log p\left(x, \hat{\theta}\right) dx - \int p_0(x) \log p(x, \theta) dx \\
=&\int p_0(x) \left\{ \sum_{k=1}^{K} \hat{\theta}_k x^k - \log \left[\int \exp \left(\sum_{k=1}^{K} \hat{\theta}_k v^k \right) dv \right] \right\} dx \\
&- \int p_0(x) \left\{ \sum_{k=1}^{K} \theta_k x^k - \log \left[\int \exp \left(\sum_{k=1}^{K} \theta_k v^k \right) dv \right] \right\} dx \\
=&\int p\left(x, \hat{\theta}\right) \left\{ \sum_{k=1}^{K} \hat{\theta}_k x^k - \log \left[\int \exp \left(\sum_{k=1}^{K} \hat{\theta}_k v^k \right) dv \right] \right\} dx \\
&- \int p\left(x, \hat{\theta}\right) \left\{ \sum_{k=1}^{K} \theta_k x^k - \log \left[\int \exp \left(\sum_{k=1}^{K} \theta_k v^k \right) dv \right] \right\} dx \\
=&\int p\left(x, \hat{\theta}\right) \log p\left(x, \hat{\theta}\right) dx - \int p\left(x, \hat{\theta}\right) \log p(x, \theta) dx \\
=&D\left(p\left(x, \hat{\theta}\right) : p(x, \theta)\right).
\end{aligned}
$$

The third equality holds because $p\left(x, \hat{\theta}\right)$ shares the same first $2K$ moments with $p_0(x)$. We then have the Pythagorean-like identity

$$
D\left(p_0(x) : p(x, \theta)\right) = D\left(p_0(x) : p\left(x, \hat{\theta}\right)\right) + D\left(p\left(x, \hat{\theta}\right) : p(x, \theta)\right).
$$

Since $D\left(p\left(x, \hat{\theta}\right) : p(x, \theta)\right) \geq 0$, it follows that

$$
D\left(p_0(x) : p\left(x, \hat{\theta}\right)\right) \leq D\left(p_0(x) : p(x, \theta)\right),
$$

where the equality holds if and only if $p\left(x, \hat{\theta}\right) = p(x, \theta)$ everywhere. □

Proof of Theorem 6. Under Assumption 3.1 to 3.7, we have $\tilde{\mu}_k \xrightarrow{as} \mu_k$ for $k = 1, \ldots, K$. The proof then follows exactly as that of Theorem 3. □

Proof of Collorary 7. The proof is identical to that of Collorary 4. □

References

Abadie, A, J. Angrist and G. Imbens. "Instrumental Variables Estimation of Quantile Treatment Effects." *Econometrica*, 70 (2002): 91-117.

Athey, S., and G.W. Imbens. "Identification and Inference in Nonlinear Difference-in-Difference Models." *Econometrica,* 74 (2006): 431-497.

Barron, A.R., and C. Sheu. "Approximation of Density Functions by Sequences of Exponential Families." *Annals of Statistics* 19 (1991): 1347-69.

Biddle, J., L. Boden and R. Reville. "A Method for Estimating the Full Distribution of a Treatment Effect, with an Application to the Impact of Workplace Injury on Subsequent Earnings." Working Paper, 2003.

Csiszár, I. "I-divergency Geometry of Probability Distributions and Minimization Problems." *Annals of Probability,* 3 (1975): 146-158.

Carroll, R. and P. Hall. "Optimal Rates of Convergence for Deconvolving a Density." *Journal of American Statistical Association* 83 (1988): 1184-1186.

Carroll, R., and P. Hall. "Low-Order Approximations in Deconvolution and Regression with Errors in Variables." *Journal of Royal Statistical Society. B* 66 (2004): 31-46.

Chernozhukov, V. and C. Hansen. "An IV Model of Quantile Treatment Effects." *Econometrica*, 73 (2005): 245-261.

Cobb, L., P. Koppstein, and N. H. Chen. "Estimation and moment recursion relations for multimodal distributions of the exponential family." *Journal of the American Statistical Association* 78 (1983): 124-130.

Dalén, J. "Algebraic Bounds on Standardized Sample Moments." *Statistics and Probability Letters* 5 (1987): 329-31.

Fan, J. "Deconvolution with Supersmooth Distributions." *Canadian Journal of Statistics* 20 (1992): 155-169.

Firpo, S. "Efficient Semiparametric Estimation of Quantile Treatment Effects." Working

Paper, University of British Columbia, 2004.

Golan, A., G. Judge, and D. Miller. Maximum Entropy Econometrics: Robust Estimation with Limited Data. New York: John Wiley and Sons, 1996.

Heckman, J. and J. Smith. "Evaluating the Welfare State." In *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centenary*, edited by S. Strom. Cambridge: Cambridge University Press, 1998.

Heckman, J., J. Smith and N. Clements. "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies 64*, no. 4 (1997): 487-535.

Heckman, J. and E. Vytlacil "Structural Equations, Treatment Effects and Econometric Policy Evaluation." *Econometrica*, 73(2005): 669-738.

Horowitz, J. and M. Markatou. "Semiparametric Estimation of Regression Models for Panel Data." *Review of Economic Studies* 63 (1996): 145-68.

Imbens, G.W. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." *Review of Economics and Statistics* 86 (2004): 4-29.

Jaynes, E. T. "Information Theory and Statistical Mechanics." *Physics Review* 106 (1957): 620-30.

White, H. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50 (1982): 1-26.

White, H. and G. M. McDonald. "Some Large-sample Tests for Nonnormality in the Linear Regression Model." *Journal of American Statistical Association* 75 (1982): 16-28.

Wu, X. "Calculation of Maximum Entropy Densities with Application to Income Distribution." *Journal of Econometrics* 115 (2003): 347-54.

Zellner, A. and R. A. Highfield. "Calculation of Maximum Entropy Distribution and Approximation of Marginal Posterior Distributions." *Journal of Econometrics* 37 (1988): 195-209.

Table 1. Simulation results

			Maxent Density		Orthogonal Series	
N	Δ	ε	(a)	(b)	(a)	(b)
250	Normal	Normal	0.002	0.011	0.001	0.006
250	Normal	Uniform	0.002	0.011	0.001	0.006
250	Skew-Normal	Normal	0.042	0.643	0.051	0.855
250	Skew-Normal	Uniform	0.042	0.644	0.050	0.837
250	Mix-Normal	Normal	0.029	0.058	0.117	0.688
250	Mix-Normal	Uniform	0.023	0.058	0.118	0.704
500	Normal	Normal	0.001	0.005	0.001	0.003
500	Normal	Uniform	0.001	0.006	0.001	0.003
500	Skew-Normal	Normal	0.041	0.601	0.051	0.861
500	Skew-Normal	Uniform	0.041	0.615	0.051	0.859
500	Mix-Normal	Normal	0.023	0.048	0.117	0.695
500	Mix-Normal	Uniform	0.019	0.048	0.118	0.698

(a): Integrated squared errors
(b): Mean squared errors of inter-quartile range

Table 2. Estimated moments of the treatment distribution (unit: $1,000)

	μ_1	μ_2	μ_3	μ_4	s.e.
Unconditional	0.77	8.75	185.86	6280.52	2.86
Conditional	0.84	8.61	173.20	7609.69	2.81
Interaction	0.75	8.51	176.39	7712.65	2.82

μ_i: the i^{th} moment of individual treatment effect distribution, i=1,2,3,4
s.e.: the standard error of individual treatment effect distribution

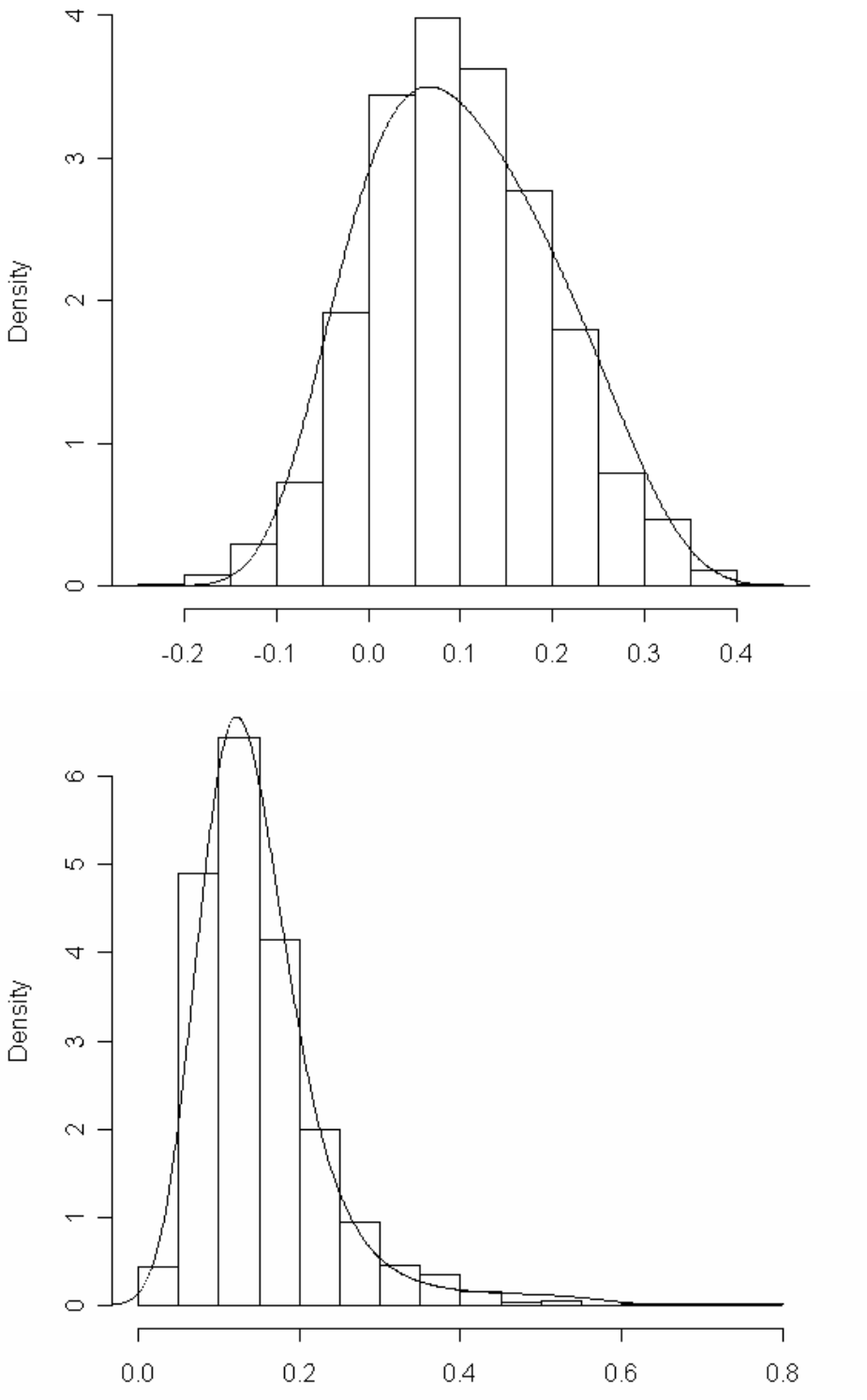

Figure 1: Estimation of random treatment effect distribution

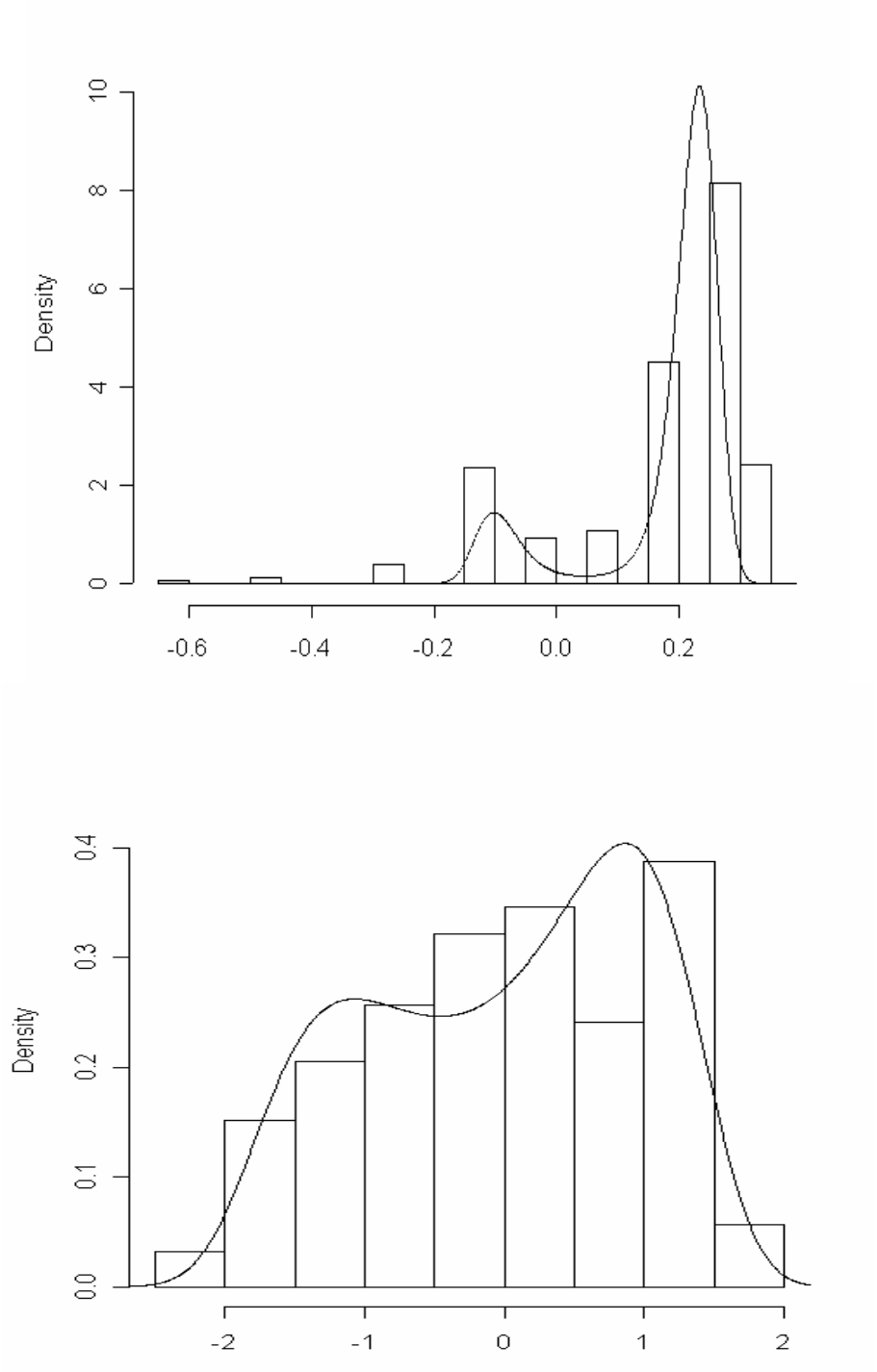

Figure 2. Estimation of heterogenous treatment effect distribution

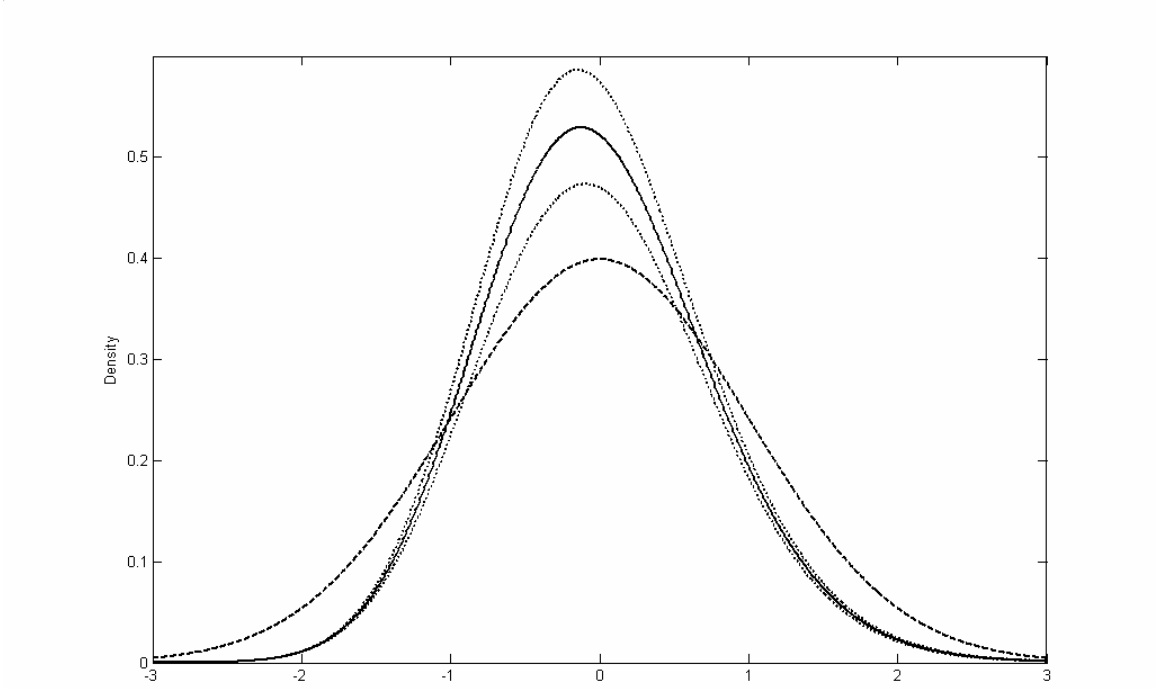

Figure 3: Estimated treatment effect distribution (solid) with asymptotic standard errors (dotted), and normal distribution with identical mean and variance (dashed); unit: $1,000