ECONOMETRICA

JOURNAL OF THE ECONOMETRIC SOCIETY

An International Society for the Advancement of Economic Theory in its Relation to Statistics and Mathematics

http://www.econometricsociety.org/

Econometrica, Vol. 81, No. 3 (May, 2013), 1185–1201

ROBUSTNESS, INFINITESIMAL NEIGHBORHOODS, AND MOMENT RESTRICTIONS

YUICHI KITAMURA
Cowles Foundation for Research in Economics, Yale University, New Haven, CT 06520, U.S.A.

TAISUKE OTSU
London School of Economics and Political Science, London, WC2A 2AE, U.K.

KIRILL EVDOKIMOV
Princeton University, Princeton, NJ 08544, U.S.A.

Econometrica, Vol. 81, No. 3 (May, 2013), 1185–1201

NOTES AND COMMENTS

ROBUSTNESS, INFINITESIMAL NEIGHBORHOODS, AND MOMENT RESTRICTIONS

BY YUICHI KITAMURA, TAISUKE OTSU, AND KIRILL EVDOKIMOV[1]

This paper is concerned with robust estimation under moment restrictions. A moment restriction model is semiparametric and distribution-free; therefore it imposes mild assumptions. Yet it is reasonable to expect that the probability law of observations may have some deviations from the ideal distribution being modeled, due to various factors such as measurement errors. It is then sensible to seek an estimation procedure that is robust against slight perturbation in the probability measure that generates observations. This paper considers local deviations within shrinking topological neighborhoods to develop its large sample theory, so that both bias and variance matter asymptotically. The main result shows that there exists a computationally convenient estimator that achieves optimal minimax robust properties. It is semiparametrically efficient when the model assumption holds, and, at the same time, it enjoys desirable robust properties when it does not.

KEYWORDS: Asymptotic Minimax Theorem, Hellinger distance, semiparametric efficiency.

1. INTRODUCTION

CONSIDER A PROBABILITY MEASURE $P_0 \in \mathcal{M}$, where $\mathcal{M}$ is the set of all probability measures on the Borel σ-field $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ of $\mathcal{X} \subseteq \mathbb{R}^d$. Let $g: \mathcal{X} \times \Theta \to \mathbb{R}^m$ be a vector of functions parameterized by a p-dimensional vector θ which resides in $\Theta \subset \mathbb{R}^p$. The function g satisfies

$$E_{P_0}[g(x, \theta_0)] = \int g(x, \theta_0)\, dP_0 = 0, \quad \theta_0 \in \Theta. \tag{1.1}$$

The moment condition model (1.1) is semiparametric and distribution-free; therefore it imposes mild assumptions. Nevertheless, it is reasonable to expect that the probability law of observations may have some deviations from the restriction under the moment condition model. It is then sensible to seek for estimation and testing procedures that are robust against slight perturbations in the observed data, or more formally, perturbations in the probability measure that generates observations. This notion of robustness can be illustrated as follows. Let a functional $\theta(P)$, $P \in \mathcal{M}$ solve the moment condition model (1.1), in

[1]We are grateful to a co-editor and two anonymous referees for helpful remarks and suggestions. We thank participants at the CIREQ Conference on GMM, the 2007 Winter Meetings of the Econometric Society, the 2007 Netherlands Econometrics Study Group Annual Conference, and seminars at Boston University, Chicago Booth, Harvard-MIT, LSE, NYU, Ohio State, Seoul National University, Texas A&M, the University of Tokyo, Vanderbilt, and Wisconsin for valuable comments. We acknowledge financial support from the National Science Foundation via Grants SES-0241770, SES-0551271, and SES-0851759 (Kitamura) and SES-0720961 (Otsu).

 DOI: 10.3982/ECTA8617

the sense that $\theta_0 = \theta(P_0)$. Suppose, however, that observations $x_1, \dots, x_n$ are not drawn according to P_0, but according to its "perturbed" version P instead. This can be attributed to various factors, including measurement errors or data contamination. These are imminent and realistic concerns in applications. The goal of robust estimation is to obtain an estimator $\bar{\theta} = \bar{\theta}(x_1, \dots, x_n)$ that is not sensitive to such perturbations, so that the deviation of the estimated value $\bar{\theta}$ from the parameter value of interest $\theta_0 = \theta(P_0)$ remains stable. Decompose the deviation as

$$\bar{\theta} - \theta_0 = \left[\bar{\theta} - \theta(P)\right] + \left[\theta(P) - \theta(P_0)\right]. \tag{1.2}$$

In the asymptotic mean squared error (MSE) calculation presented below, the expectation of the square of the term in the first square bracket contributes to the variance of the estimator, whereas the second corresponds to the bias. An estimator that achieves small MSE *uniformly* in P over a neighborhood of P_0 is desirable.

Asymptotic theory of robust estimation when the model is parametric has been considered extensively in the literature; see Rieder (1994) for a comprehensive survey. In a pioneering paper, Beran (1977) discussed "robust and efficient" estimation of parametric models. Suppose P_θ, $\theta \in \Theta \subset \mathbb{R}^k$ is a parametric family of probability measures. Observations are drawn from a probability law P, which may not be a member of the parametric family. Let p_θ and p denote the densities associated with the probability measures P_θ and P. It is well known that the parametric MLE procedure corresponds to minimizing the objective function $\rho = \int \log(p/p_\theta)\, p\, dx$. Beran pointed out that a small change in the density p can lead to a large change in the objective function ρ (note the log in ρ), implying the nonrobustness of the MLE. He showed that the parametric minimum Hellinger distance estimator (MHDE) is "robust and efficient," in the sense that (i) it has an asymptotic minimax robust property, and (ii) it is asymptotically efficient when the model assumption is satisfied, that is, when the sample is generated from $P_0 = P_{\theta_0}$, where θ_0 is the true value of the parameter of interest. Let $H(P_\theta, P) = \sqrt{\int (p_\theta^{1/2}(x) - p^{1/2}(x))^2\, dx}$ denote the Hellinger distance between P_θ and P (a slightly more general definition of the Hellinger distance is given in the next section). The MHDE for the parametric model is

$$\begin{aligned}\hat{\theta} &= \operatorname*{argmin}_\theta H(P_\theta, \hat{P}) \\ &= \operatorname*{argmin}_\theta \int \left(p_\theta^{1/2}(x) - \hat{p}^{1/2}(x)\right)^2 dx,\end{aligned}$$

where $\hat{p}$ is a nonparametric density estimator, such as a kernel density estimator, for P, and $\hat{P}$ is the corresponding estimator for the probability measure of x. The MHDE is asymptotically equivalent to MLE and thus efficient if the model assumption is satisfied. One can replace the Hellinger distance with other divergence measures such as the Kolmogorov–Smirnov distance, which would make the corresponding minimum divergence estimator even more robust, but it would incur efficiency loss. The parametric MHDE has been studied extensively and applied to various models.

The parametric MHDE has theoretical advantages and excellent finite sample performance documented by numerous simulation studies, but it has limitations as well. It requires the nonparametric density estimator when at least some components of x are continuously distributed. This makes its practical application inconvenient, and is problematic when x is high dimensional, due to the curse of dimensionality. It also necessitates the evaluation of the integral $\int (p_\theta^{1/2}(x) - \hat{p}^{1/2}(x))^2\,dx$. This involves either numerical integration or an approximation by an empirical average with inverse density weighting using a nonparametric density estimator. The former can be hard to implement, and the latter may have undesirable effects in finite samples. This paper aims at developing robust methods for moment restriction models, by applying the MHDE procedure. The resulting estimator is semiparametrically efficient when the model assumption holds, and, at the same time, it enjoys an optimal minimax robust property when it does not. The implementation of the estimator is easy. Unlike its parametric predecessor, it requires neither nonparametric density estimation nor evaluation of integration.

2. PRELIMINARIES

The econometrician wishes to estimate the unknown θ_0 in (1.1). Suppose a random sample $\{x_i\}_{i=1}^n$ generated from P is observed. As discussed in Section 1, our focus is on robust estimation of θ_0 when the probability measure P, from which the observations are drawn, is a (locally) perturbed version of P_0, not P_0 itself. There exists an extensive literature concerning the estimation of (1.1) under the "classical" setting where data are indeed drawn from P_0. Many estimators for θ_0 are available, including GMM (Hansen (1982)), the empirical likelihood (EL) estimator, and its variants. This paper is concerned with an estimator that can be viewed as MHDE applied to the moment restriction model (1.1). The Hellinger distance between two probability measures is defined as follows:

DEFINITION 2.1: Let P and Q be probability measures with densities p and q with respect to a dominating measure ν. The Hellinger distance between P and Q is then given by

$$H(P,Q) = \left\{\int \left(p^{1/2} - q^{1/2}\right)^2 d\nu\right\}^{1/2} = \left\{2 - 2\int p^{1/2} q^{1/2}\,d\nu\right\}^{1/2}.$$

It is often convenient to use the standard notation in the literature that does not explicitly refer to the dominating measure. Then the above definition becomes

$$H(P,Q)=\left\{\int \left(dP^{1/2}-dQ^{1/2}\right)^2\right\}^{1/2}=\left\{2-2\int dP^{1/2}\,dQ^{1/2}\right\}^{1/2}.$$

Here we show some results concerning the Hellinger distance that are useful in understanding the robustness theorems in the next section.

DEFINITION 2.2: Let P and Q be probability measures with densities p and q with respect to a dominating measure ν. The α-divergence from Q to P is given by

$$I_\alpha(P,Q)=\frac{1}{\alpha(1-\alpha)}\int\left(1-\left(\frac{p}{q}\right)^\alpha\right)q\,d\nu,\quad \alpha\in\mathbb{R}.$$

If P is not absolutely continuous with respect to Q, then $\int \mathbb{I}\{p>0,q=0\}\,d\nu>0$, and as a consequence, $I_\alpha(P,Q)=\infty$ for $\alpha\geq 1$. A similar argument shows that $I_\alpha(P,Q)=\infty$ if $Q\not\ll P$ and $\alpha\leq 0$. Note that I_α is well defined for $\alpha=0$ by taking the limit $\alpha\to 0$ in the definition. Indeed, L'Hospital's rule implies that

$$\lim_{\alpha\to 0} I_\alpha(P,Q)=\int \log\left(\frac{p}{q}\right)q\,d\nu := K(P,Q)$$

(with the above convention for the case where $P\not\ll Q$), giving rise to the well-known Kullback–Leibler (KL) divergence measure from Q to P. The case with $\alpha=1$ corresponds to the KL divergence with the roles of P and Q reversed. Note that the above definitions imply that the α-divergence includes the Hellinger distance as a special case, in the sense that

$$H^2(P,Q)=\frac{1}{2}I_{1/2}(P,Q).$$

The proofs of all the results are given in the Supplemental Material (Kitamura, Otsu and Evdokimov (2013)).

LEMMA 2.1: *For probability measures P and Q,*

$$\max(\alpha,1-\alpha)I_\alpha(P,Q)\geq\frac{1}{2}I_{1/2}(P,Q)$$

for every $\alpha\in\mathbb{R}$.

REMARK 2.1: Lemma 2.1 has some implications on a neighborhood system generated by the Hellinger distance. Consider the following neighborhood of a probability measure P whose radius in terms of I_α is $\delta > 0$:

$$B_{I_\alpha}(P, \delta) = \{Q \in \mathcal{M} : \sqrt{I_\alpha(Q, P)} \leq \delta\}.$$

Lemma 2.1 implies that

$$I_\alpha(P, Q) \geq \frac{1}{2\left(\left(\frac{1}{2} + L\right) \vee \left(\frac{1}{2} + U\right)\right)} I_{\alpha_0}(P, Q)$$

holds for every $\alpha \in [\frac{1}{2} - L, \frac{1}{2} + U]$, where $L, U > 0$ determine the lower and upper bounds for the range of α, if $\alpha_0 = \frac{1}{2}$. It is easy to verify that this statement holds only if $\alpha_0 = \frac{1}{2}$. Now, define

$$C(L, U) = \left(\frac{1}{2} + L\right) \vee \left(\frac{1}{2} + U\right);$$

then by the above inequality,

$$\bigcup_{\alpha \in [1/2 - L, 1/2 + U]} B_{I_\alpha}(P_0, \delta) \subset B_{I_{1/2}}\big(P_0, \sqrt{2C(L, U)}\delta\big). \tag{2.1}$$

That is, the union of the I_α-based neighborhoods over $\alpha \in [\frac{1}{2} - L, \frac{1}{2} + U]$ is covered by the Hellinger neighborhood $B_{I_{1/2}}$ with a “margin” given by the multiplicative constant $\sqrt{2C(L, U)}$. Equation (2.1) is important, since in what follows we consider robustness of estimators against perturbation of P_0 within its neighborhood, and it is desirable to use a neighborhood that is sufficiently large to accommodate a large class of perturbations. The inclusion relationship shows that the Hellinger-based neighborhood covers other neighborhood systems based on I_α, $\alpha \in [\frac{1}{2} - L, \frac{1}{2} + U]$ if the radii are chosen appropriately. It is easy to verify that (2.1) does not hold if the Hellinger distance $I_{1/2}$ is replaced by I_α, $\alpha \neq \frac{1}{2}$, showing the special status of the Hellinger distance among the α-divergence family.

REMARK 2.2: Lemma 2.1 is a statement for every pair of measures (P, Q); thus it holds even if $P \not\ll Q$ or $Q \not\ll P$. On the other hand, it is useful to consider the behavior of I_α when one of the two measures is not absolutely continuous with respect to the other. Consider a sequence of probability measures $\{P^{(n)}\}_{n \in \mathbb{N}}$. Suppose $I_\alpha(P^{(n)}, P_0) \to 0$ for an $\alpha \in \mathbb{R}$; then $I_{\alpha'}(P^{(n)}, P_0) \to 0$ for every $\alpha' \in (0, 1)$. But the reverse (i.e., reversing the roles of α and α') is not true. If $P^{(n)}$, $n \in \mathbb{N}$ are not absolutely continuous with respect to P_0, $I_{\alpha'}(P^{(n)}, P_0) = \infty$ for every $\alpha' \geq 1$ even if $\rho_\alpha(P^{(n)}, P_0) \to 0$ for $\alpha \in (0, 1)$ (and a similar argument holds for $\alpha' \leq 0$). This shows that I_α-based neighborhoods with $\alpha \notin (0, 1)$ are too small: there are measures that are outside of $B_{I_\alpha}(P_0, \delta)$, $\alpha \notin (0, 1)$ no matter how large δ is, or how close they are to P_0 in terms of, say, the Hellinger distance H.

REMARK 2.3: The inequality in Lemma 2.1 might be of interest on its own, as it generalizes many inequalities in the literature. For $\alpha = 1$ or 0, it corresponds to the well-known inequality between the KL divergence and the Hellinger distance

$$H(P, Q)^2 \leq K(P, Q); \tag{2.2}$$

see, for example, Pollard (2002, p. 62). Another commonly used definition of divergence between probability measures is the χ^2 distance. It is given, if $P \ll Q \ll \nu$, by $\chi^2(P, Q) = \int \frac{(p-q)^2}{q}\, d\nu$, and it is shown that

$$H(P, Q)^2 \leq \chi^2(P, Q) \tag{2.3}$$

(Reiss (1989)). This is implied by Lemma 2.1 by letting $\alpha = 2$. Proposition 3.1 in Zhang (2006) is closer to our result in terms of its generality; it shows that $\max(\alpha, 1-\alpha) I_\alpha(P, Q) \geq \frac{1}{2} I_{1/2}(P, Q)$ holds for $\alpha \in [0, 1]$, which covers (2.2) but not (2.3)[2]. Lemma 2.1 shows that this type of inequality holds for all $\alpha \in \mathbb{R}$.

Beran (1977), considering a parametric model, proposed MHDE that minimizes the Hellinger distance between a model-based probability measure (from the parametric family) and a nonparametric probability measure estimate. An application of the MHDE procedure to the moment condition model (1.1) yields a computationally simple procedure as follows. Let P_n denote the empirical measure of observations $\{x_i\}_{i=1}^n$. P_n is an appropriate model-free estimator in our construction of the MHDE. Let

$$\mathcal{P}_\theta = \left\{ P \in \mathcal{M} : \int g(x, \theta)\, dP = 0 \right\}$$

and

$$\mathcal{P} = \bigcup_{\theta \in \Theta} \mathcal{P}_\theta; \tag{2.4}$$

then the MHDE, denoted by $\hat{\theta}$, is defined to be a parameter value that solves the optimization problem

$$\inf_{\theta \in \Theta} \inf_{P \in \mathcal{P}_\theta} H(P, P_n) = \inf_{P \in \mathcal{P}} H(P, P_n).$$

By convex duality theory (Kitamura (2006)), the objective function has the following representation:

$$\inf_{P \in \mathcal{P}_\theta} H(P, P_n) = \max_{\gamma \in \mathbb{R}^m} -\frac{1}{n} \sum_{i=1}^n \frac{1}{1 + \gamma' g(x_i, \theta)}.$$

[2]Zhang (2006) also derived a lower bound for the Hellinger distance in terms of I_α.

Therefore the MHDE is $\hat{\theta} = \arg\min_{\theta\in\Theta} \max_{\gamma\in\mathbb{R}^m} -\frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\gamma' g(x_i,\theta)}$, which is easy to compute.

It is easy to verify that we can obtain the MHDE as a Generalized Empirical Likelihood (GEL) estimator by letting $\gamma = -1/2$ in equation (2.6) of Newey and Smith (2004). Asymptotic properties of the (G)EL estimators for θ_0 in (1.1), when data drawn from P_0 are observed, are well understood (see, e.g., Kitamura and Stutzer (1997), Smith (1997), Imbens, Spady, and Johnson (1998), Newey and Smith (2004)). Let $G = E_{P_0}[\partial g(x,\theta_0)/\partial\theta']$, $\Omega = E_{P_0}[g(x,\theta_0)g(x,\theta_0)']$, and $\Sigma = G'\Omega^{-1}G$. Then

$$\sqrt{n}(\hat{\theta}_{\mathrm{GEL}} - \theta_0) \xrightarrow{d} N\big(0, \Sigma^{-1}\big). \tag{2.5}$$

It follows that the MHDE and other GEL estimators are semiparametrically efficient in the absence of data perturbation. At the same time, the MHDE possesses a distinct property of being asymptotic optimal robust if observations are drawn from a perturbed version of P_0, as we shall see in the next section.

3. ROBUST ESTIMATION THEORY

We now analyze robustness of the MHDE $\hat{\theta}$. Define a functional

$$T(P) = \arg\min_{\theta\in\Theta}\max_{\gamma\in\mathbb{R}^m} -\int \frac{1}{1+\gamma' g(x,\theta)}\,dP;$$

then the MHDE can be interpreted as the value of functional T evaluated at the empirical measure P_n. In other words, each realization of P_n completely determines the value of the MHDE $\hat{\theta}$. To make the dependence explicit, we write $\hat{\theta} = T(P_n)$, and study properties of the mapping $T : \mathcal{M} \to \Theta$. This definition of $T(\cdot)$, however, causes a technical difficulty when the distribution of $g(x,\theta)$ is unbounded for some $\theta \in \Theta$ and $P \in \mathcal{M}$. To overcome this technical difficulty, we introduce the following mapping defined by a trimmed moment function:

$$\bar{T}(Q) = \arg\min_{\theta\in\Theta} \inf_{P\in\bar{\mathcal{P}}_\theta, P\ll Q} H(P,Q),$$

where $\{m_n\}_{n\in\mathbb{N}}$ is a sequence of positive numbers satisfying $m_n \to \infty$ as $n \to \infty$, and

$$\bar{\mathcal{P}}_\theta = \left\{P \in \mathcal{M} : \int g(x,\theta)\mathbb{I}\{x \in \mathcal{X}_n\}\,dP = 0\right\},$$

$$\mathcal{X}_n = \left\{x \in \mathcal{X} : \sup_{\theta\in\Theta}\big|g(x,\theta)\big| \le m_n\right\},$$

with the indicator function $\mathbb{I}\{\cdot\}$ and the Euclidean norm $|\cdot|$; that is, $\mathcal{X}_n$ is a trimming set to bound the moment function and $\bar{\mathcal{P}}_\theta$ is a set of probability measures satisfying the bounded moment condition $E_P[g(x,\theta)\mathbb{I}\{x \in \mathcal{X}_n\}] = 0$.

Let $\tau:\Theta \to \mathbb{R}$ be a possibly nonlinear transformation of the parameter. We first focus on the estimation problem of the transformed scalar parameter $\tau(\theta_0)$ and investigate the behavior of the bias term $\tau \circ \bar{T}(Q) - \tau(\theta_0)$ in a ($\sqrt{n}$-shrinking) Hellinger ball with radius $r > 0$ around P_0,

$$B_H(P_0, r/\sqrt{n}) = \{Q \in \mathcal{M} : H(Q, P_0) \le r/\sqrt{n}\}.$$

The transformation τ to a scalar, as used by Rieder (1994), is convenient in calculating squared biases and MSEs. One may, for example, let $\tau(\theta) = c'\theta$ using a constant p-vector c. Lemma A.1(ii) guarantees that, for each $r > 0$, the value $\bar{T}(Q)$ exists for all $Q \in B_H(P_0, r/\sqrt{n})$ and all n large enough.

ASSUMPTION 3.1: *Suppose the following conditions hold*:

(i) $\{x_i\}_{i=1}^n$ *is independent and identically distributed* (*i.i.d.*);
(ii) Θ *is compact*;
(iii) $\theta_0 \in \operatorname{int}\Theta$ *is a unique solution to* $E_{P_0}[g(x, \theta)] = 0$;
(iv) $g(x, \theta)$ *is continuous over* Θ *at each* $x \in \mathcal{X}$;
(v) $E_{P_0}[\sup_{\theta\in\Theta} |g(x, \theta)|^\eta] < \infty$ *for some* $\eta > 2$, *and there exists a neighborhood* $\mathcal{N}$ *around* θ_0 *such that* $E_{P_0}[\sup_{\theta\in\mathcal{N}} |g(x, \theta)|^4] < \infty$, $g(x, \theta)$ *is continuously differentiable a.s. in* $\mathcal{N}$, $\sup_{x\in\mathcal{X}_n,\theta\in\mathcal{N}} |\partial g(x, \theta)/\partial\theta'| = o(n^{1/2})$, *and* $E_{P_0}[\sup_{\theta\in\mathcal{N}} |\partial g(x, \theta)/\partial\theta'|^2] < \infty$;
(vi) G *has the full column rank and* Ω *is positive definite*;
(vii) $\{m_n\}_{n\in\mathbb{N}}$ *satisfies* $m_n \to \infty$, $nm_n^{-\eta} \to 0$, *and* $n^{-1/2}m_n^{1+\varepsilon} = O(1)$ *for some* $0 < \varepsilon < 2$ *as* $n \to \infty$;
(viii) τ *is continuously differentiable at* θ_0.

Assumption 3.1(i)–(vi) is standard in the literature of the GMM. Assumption 3.1(iii) is a global identification condition of the true parameter θ_0 under P_0. Assumption 3.1(iv) ensures the continuity of the mapping $\bar{T}(Q)$ in $Q \in \mathcal{M}$ for each $n \in \mathbb{N}$. Assumption 3.1(v) contains the smoothness and boundedness conditions for the moment function and its derivatives. This assumption is stronger than the one to derive the asymptotic distribution in (2.5). Assumption 3.1(vi) is a local identification condition for θ_0. This assumption guarantees that the asymptotic variance matrix Σ^{-1} exists. Assumption 3.1(vii) is on the trimming parameter m_n. If $m_n \sim n^a$, this assumption is satisfied for $1/\eta < a < 1/2$. Assumption 3.1(viii) is a standard requirement for the parameter transformation τ. To characterize a class of estimators to be compared with the MHDE, we introduce the following definition.

DEFINITION 3.1: Let $T_a(P_n)$ be an estimator of θ_0 based on a mapping $T_a:\mathcal{M} \to \Theta$. Also, let $P_{\theta,\zeta}$ be a regular parametric submodel (see Bickel, Klassen, Ritov, and Wellner (1993, p. 12) or Newey (1990)) of $\mathcal{P}$ in (2.4) such that $P_{\theta_0,0} = P_0$ and $P_{\theta_0+t/\sqrt{n},\zeta_n} \in B_H(P_0, r/\sqrt{n})$ holds for $\zeta_n = O(n^{-1/2})$ eventually.

(i) T_a is called *Fisher consistent* if, for every $\{P_{\theta_n,\zeta_n}\}_{n\in\mathbb{N}}$ and $t\in\mathbb{R}^p$,

$$\sqrt{n}\big(T_a(P_{\theta_0+t/\sqrt{n},\zeta_n})-\theta_0\big)\to t. \tag{3.1}$$

(ii) T_a is called *regular* for θ_0 if, for every $\{P_{\theta_n,\zeta_n}\}_{n\in\mathbb{N}}$ with $(\theta_n',\zeta_n')'=(\theta_0',0)'+O(n^{-1/2})$, there exists a probability measure M such that

$$\sqrt{n}\big(T_a(P_n)-T_a(P_{\theta_n,\zeta_n})\big)\xrightarrow{d} M \quad \text{under} \quad P_{\theta_n,\zeta_n}, \tag{3.2}$$

where the measure M does not depend on the sequence $(\theta_n',\zeta_n')'$.

Both conditions are weak and satisfied by GMM, (G)EL, and other standard estimators. For example, the mapping T_a for the continuous updating GMM estimator (CUE) is given by

$$T_{\mathrm{CUE}}(P)=\operatorname*{argmin}_{\theta\in\Theta}\left[\int g(x,\theta)\,dP\right]'\left[\int g(x,\theta)g(x,\theta)\,dP\right]^{-1} \times\left[\int g(x,\theta)\,dP\right],$$

and, under Assumption 3.1, $T_{\mathrm{CUE}}(P_{\theta_0+t/\sqrt{n}},\zeta_n)=\theta_0+t/\sqrt{n}$ for large n. CUE therefore trivially satisfies (3.1). The regularity condition (3.2) is standard in the literature of semiparametric efficiency; see, for example, Bickel et al. (1993).

The following theorem shows the optimal robustness of the (trimmed) MHDE in terms of its maximum bias.

THEOREM 3.1: *Suppose that Assumption* 3.1 *holds*.

(i) *For every T_a that is Fisher consistent*,

$$\liminf_{n\to\infty}\sup_{Q\in B_H(P_0,r/\sqrt{n})} n\big(\tau\circ T_a(Q)-\tau(\theta_0)\big)^2\geq 4r^2B^*,$$

for each $r>0$, where $B^*=(\frac{\partial\tau(\theta_0)}{\partial\theta})'\Sigma^{-1}(\frac{\partial\tau(\theta_0)}{\partial\theta})$.

(ii) *The mapping $\bar{T}$ is Fisher consistent and satisfies*

$$\lim_{n\to\infty}\sup_{Q\in B_H(P_0,r/\sqrt{n})} n\big(\tau\circ\bar{T}(Q)-\tau(\theta_0)\big)^2=4r^2B^*,$$

for each $r>0$.

REMARK 3.1: The above result is concerned with deterministic properties of T_a and T. $T_a(Q)$ and $T(Q)$ can be regarded as the (probability) limit of the estimators $T_a(P_n)$ and $T(P_n)$ under Q, and therefore the terms evaluated here correspond to the bias of each estimator due to the deviation of Q from P_0. The theorem says that, in the class of all mappings that are Fisher consistent, the mapping $\bar{T}$ has the smallest maximum bias over the set $B_H(P_0, r/\sqrt{n})$. The (trimmed version of) the Hellinger-based mapping $\bar{T}$ is therefore optimally robust asymptotically in a minimax sense. The term $4r^2B^*$ provides a sharp lower bound for maximum squared bias, and it is attained by $\bar{T}$.

REMARK 3.2: The theorem is concerned with the trimmed version of the MHDE. It avoids the complications associated with the existence of $T(Q)$ for certain Q's. If the support of $\sup_{\theta\in\Theta}|g(x,\theta)|$ is bounded under every $Q \in B_H(P_0, r/\sqrt{n})$ for large enough n (e.g., if the moment function g is bounded), then we do not need the trimming term $\mathbb{I}\{x \in \mathcal{X}_n\}$. In this case, the mapping T without trimming has the above optimal robust property.

REMARK 3.3: The index n in the statement of Theorem 3.1 simply parameterizes how close $Q \in B_H(P_0, r/\sqrt{n})$ and P_0 are, and does not have to be interpreted as the sample size. The next theorem, however, is concerned with MSEs and the index n represents the sample size there.

The next theorem is our main result, which is concerned with (the supremum of) the MSE of the minimum Hellinger distance estimator $\hat{\theta} = T(P_n)$ and other competing estimators. Let

$$\bar{B}_H(P_0, r/\sqrt{n}) = B_H(P_0, r/\sqrt{n}) \cap \left\{Q \in \mathcal{M} : E_Q\left[\sup_{\theta\in\Theta}\left|g(x,\theta)\right|^{\eta}\right] < \infty\right\}. \tag{3.3}$$

We use the notation $P^{\otimes n}$ to denote the n-fold product measure of a probability measure P.

THEOREM 3.2: *Suppose that Assumption 3.1 holds.*
(i) *For every Fisher consistent and regular mapping T_a,*

$$\lim_{b\to\infty}\liminf_{n\to\infty}\sup_{Q\in\bar{B}_H(P_0,r/\sqrt{n})}\int b \wedge n\big(\tau\circ T_a(P_n) - \tau(\theta_0)\big)^2\,dQ^{\otimes n} \geq \big(1+4r^2\big)B^*,$$

for each $r > 0$.
(ii) *The mapping T is Fisher consistent and regular, and the MHDE $\hat{\theta} = T(P_n)$ satisfies*

$$\lim_{b\to\infty}\lim_{n\to\infty}\sup_{Q\in\bar{B}_H(P_0,r/\sqrt{n})}\int b \wedge n\big(\tau\circ T(P_n) - \tau(\theta_0)\big)^2\,dQ^{\otimes n} = \big(1+4r^2\big)B^*,$$

for each $r > 0$.

REMARK 3.4: This theorem establishes an asymptotic minimax optimality property of the MHDE, in terms of MSE among all the estimators, that satisfies the two conditions in Definition 3.1. Note that the expression $\sup_{Q\in\bar{B}_H(P_0,r/\sqrt{n})}\int b\wedge n(\tau\circ T_a(P_n)-\tau(\theta_0))^2\,dQ^{\otimes n}$ is the maximum finite sample MSE of $T_a(P_n)$. Thus our criterion for evaluating T_a (and T) is the limit of its maximum finite sample MSE. Taking the supremum over B_H before letting n go to infinity is important for capturing finite sample robustness properties. The method of calculating the truncated MSE first, then letting $b\to\infty$, is standard in the literature of robust estimation, but is also used in general contexts; see, for example, Bickel (1981) and LeCam and Yang (1990). Once again, we are able to derive a sharp lower bound for the maximum MSE and show that it is achieved by the MHDE $\hat{\theta}=T(P_n)$.

REMARK 3.5: Unlike in Theorem 3.1, optimality is achieved by the untrimmed version of the MHDE. Note that $T(P_n)$ exists for large n under Assumption 3.1, in contrast to our discussion in Remark 3.2 on Theorem 3.1. Theorem 3.2, however, restricts the robustness neighborhood by an extra requirement as in (3.3). This is useful in showing that the untrimmed MHDE achieves the lower bound.

REMARK 3.6: Theorem 3.2 proves that the MHDE is asymptotically optimally robust over a sequence of infinitesimal neighborhoods. Note that the Hellinger neighborhood over which the maximum of MSE is taken is nonparametric, in the sense that potential deviations from P_0 cannot be indexed by a finite dimensional parameter. That is, our robustness concept demands uniform robustness over a nonparametric, infinitesimal neighborhood. The use of infinitesimal neighborhoods, where the radius of the Hellinger ball shrinks at the rate $n^{1/2}$, is useful in balancing the magnitude of bias and variance in our asymptotics. If one uses a fixed, global neighborhood, then the bias term would dominate the behavior of estimators. This may fail to provide a good approximation of finite sample behavior in actual applications, since in reality it would be reasonable to be concerned with both the stochastic fluctuation of estimators and their deterministic bias due to, say, data contamination. We note that there is a related but distinct literature on the asymptotics theory when the model is globally misspecified, as in White (1982), who considered parametric MLE. Kitamura (1998, 2002) offered such analysis for conditional and unconditional moment condition models. Moreover, Schennach (2007) provided novel and potentially very useful results of EL estimators and their variants in misspecified moment condition models. We regard our paper as a complement to, rather than a substitute for, the results obtained in these papers. There are fundamental differences between the characteristics of the problems the current paper considers and those of the papers on misspecification. First, our object of interest is θ_0, not a pseudo-true value, as we consider data perturbation rather than model misspecification. Second, the nature of our analysis is local and, therefore, the parameter value θ_0 in (1.1) is still identified asymptotically. Third, as noted above, we consider uniform robustness over a nonparametric neighborhood. The papers cited above consider pointwise problems. Therefore our approach deals with phenomena that are very different from the ones analyzed in the literature of misspecified models.

REMARK 3.7: We have seen in Remark 2.1 that the Hellinger neighborhood B_H has nice and distinct properties, in particular the inclusion relationship (2.1). The Hellinger neighborhood B_H is commonly used in the literature of robust estimation (of parametric models); see, for example, Beran (1977), Bickel (1981), and Rieder (1994). We should note, however, that other neighborhood systems have been used in the literature as well. For example, one may replace the Hellinger distance H with the Kolmogorov–Smirnov (KS) distance in the definition of B_H. As Beran (1984) noted, however, to guarantee robustness in the Kolmogorov–Smirnov neighborhood system, one needs

> "to use minimum distance estimates based on the Kolmogorov–Smirnov metric or a distance weaker than the Kolmogorov–Smirnov metric ... The general principle here is that the estimation distance be no stronger than the distance describing the contamination neighborhood...."

Donoho and Liu (1988) developed a general theory of the above point. What this means is that an estimator that is robust against perturbations within Kolmogorov–Smirnov neighborhoods has to be minimizing the KS (or weaker) distance. The "minimum KS estimator" for the moment restriction model would be indeed robust, but it cannot be semiparametrically efficient when the model assumption holds. Therefore, unlike the moment restriction MHDE, the estimator is not "robust and efficient." Another drawback is its computation, since, unlike the moment restriction MHDE, no convenient algorithm to minimize the Kolmogorov–Smirnov distance under the moment restriction is known in the literature. It should be noted that the moment restriction MHDE is efficient in the sense that it achieves the semiparametric efficiency bound. It does not have the desirable higher order properties of EL (Newey and Smith (2004)) or the ETEL estimator proposed by Schennach (2007).

The above MSE theorem conveniently summarizes the desirable robustness properties of the MHDE in terms of both (deterministic) bias and variance. It has, however, some limitations. First, its minimaxity result is obtained within Fisher consistent and regular estimators. While these requirements are weak, it might be of interest to expand the class of estimators. More importantly, implicit in the MSE-based analysis is that we are interested in L^2-loss. One may wish to use other types of loss functions, however, and it is of interest to see whether the above minimax results can be extended to a larger class of loss. The next theorem addresses these two issues. Of course, the MSE has an advantage of subsuming the bias and the variance in one measure. To deal with general loss functions, the next theorem focuses on the risk of estimators around a

Fisher consistent mapping evaluated at the perturbed measure Q. This can be regarded as calculating the risk of the first bracket of the decomposition (1.2), that is, the stochastic part of the deviation of the estimator from the parameter of interest θ_0.

Let $\mathcal{S}$ be a set of all estimators, that is, the set of all $\bar{\mathbb{R}}^p$-valued measurable functions. We now investigate robust risk properties of this large class of estimators. The loss function we consider satisfies the following weak requirements.

ASSUMPTION 3.2: *The loss function* $\ell : \bar{\mathbb{R}}^p \to [0, \infty]$ *is* (i) *symmetric subconvex* (*i.e., for all* $z \in \mathbb{R}^p$ *and* $c \in \mathbb{R}$, $\ell(z) = \ell(-z)$ *and* $\{z \in \mathbb{R}^p : \ell(z) \leq c\}$ *is convex*); (ii) *upper semicontinuous at infinity*; *and* (iii) *continuous on* $\bar{\mathbb{R}}^p$.

We now present an optimal risk property for the MHDE.

THEOREM 3.3: *Suppose that Assumptions* 3.1 *and* 3.2 *hold*.

(i) *For every Fisher consistent mapping* T_a,

$$\lim_{b\to\infty}\lim_{r\to\infty}\liminf_{n\to\infty}\inf_{S_n\in\mathcal{S}}\sup_{Q\in\bar{B}_H(P_0,r/\sqrt{n})}\int b\wedge\ell\big(\sqrt{n}\big(S_n-\tau\circ T_a(Q)\big)\big)\,dQ^{\otimes n}$$
$$\geq\int\ell\,dN\big(0,B^*\big).$$

(ii) *The mapping* T *is Fisher consistent and the MHDE* $\hat{\theta} = T(P_n)$ *satisfies*

$$\lim_{b\to\infty}\lim_{n\to\infty}\sup_{Q\in\bar{B}_H(P_0,r/\sqrt{n})}\int b\wedge\ell\big(\sqrt{n}\big(\tau\circ T(P_n)-\tau\circ\bar{T}(Q)\big)\big)\,dQ^{\otimes n}$$
$$=\int\ell\,dN\big(0,B^*\big),$$

for all $r > 0$.

Note that Theorem 3.3(ii) remains valid if $T(P_n)$ is replaced by $\bar{T}(P_n)$. This theorem shows that the MHDE is once again optimally robust even for the general risk criterion, and this holds in the class of essentially all possible estimators. As noted above, the result is concerned with the stochastic component of the decomposition (1.2). Theorem 3.1 has already established that the MHDE is optimal in terms of its bias, that is, the deterministic part of the decomposition (1.2) in the second bracket. The latter result does not depend on a specific loss function. Thus the MHDE enjoys general optimal robust properties under a quite general setting, in terms of both the stochastic component and the deterministic component. Note that analyzing these two parts separately is common in the literature of robust statistics: see, for example, Rieder (1994).

4. SIMULATION

The purpose of this section is to examine the robustness properties of the MHDE and other well-known estimators such as GMM using Monte Carlo simulations. MATLAB is used for computation throughout the experiments. The sample size n is 100 for all designs, and we ran 5000 replications for each design.

The baseline simulation design in this experiment follows that of Hall and Horowitz (1996). We then "contaminate" the simulated data to explore robustness of estimators. More specifically, let $x = (x_1, x_2)' \sim N(0, 0.4^2 I_2)$. This normal law corresponds to P_0 in the preceding sections. The specification of the moment function g is

$$g(x, \theta) = \left(\exp\left\{-0.72 - \theta(x_1 + x_2) + 3x_2\right\} - 1\right)\begin{pmatrix} 1 \\ x_2 \end{pmatrix}.$$

The moment condition $\int g(x, \theta)\, dP_0 = 0$ is uniquely solved at $\theta_0 = 3$. The goal is to estimate this value using the above specification of g when the original DGP is perturbed into different directions. More specifically, we use $x \sim N(0, \Sigma_{(\delta,\rho)})$, where

$$\Sigma_{(\delta,\rho)} = 0.4^2 \begin{pmatrix} (1+\delta)^2 & \rho(1+\delta) \\ \rho(1+\delta) & 1 \end{pmatrix}.$$

The unperturbed case thus corresponds to $\delta = \rho = 0$. In the simulation, we set $\rho = 0.1\sqrt{2}\cos(2\pi\omega)$ and $\delta = 0.25\sin(2\pi\omega)$ and let ω vary over $\omega_j = j/64$, $j = 0, \dots, 63$. This yields 64 different designs; for each of them, 5000 replications are performed and RMSE and $\Pr\{|\hat{\theta} - \theta_0| > 0.5\}$ are calculated. We consider the following estimators: empirical likelihood (EL), MHDE, exponential tilting (ET), GMM (GMM2), and continuously updated GMM (CUE). GMM2 is calculated following the standard two step procedure where the initial estimate is obtained from identity weighting. CUE's performance is extremely sensitive to data perturbations considered here; its RMSE is much higher than that of the other estimators. For convenience, we only plot the results for EL, MHDE, ET, and GMM2 in displaying their RMSEs. The results are presented in Figure 1. In the left panel, each curve represents the RMSE of a particular estimator as a function of ω_j. The right panel (labeled "Pr") displays the simulated probability of an estimator deviating from the target $\theta_0 = 3$ by more than 0.5.

While RMSE is a potentially informative measure, it can be highly misleading, as some of the estimators may not have finite moments. We thus focus on the results for deviation probabilities. The performance of CUE clearly indicates its lack of robustness against data perturbations. We also see that GMM2 is affected by perturbations much more than EL, MHDE, and ET, except for

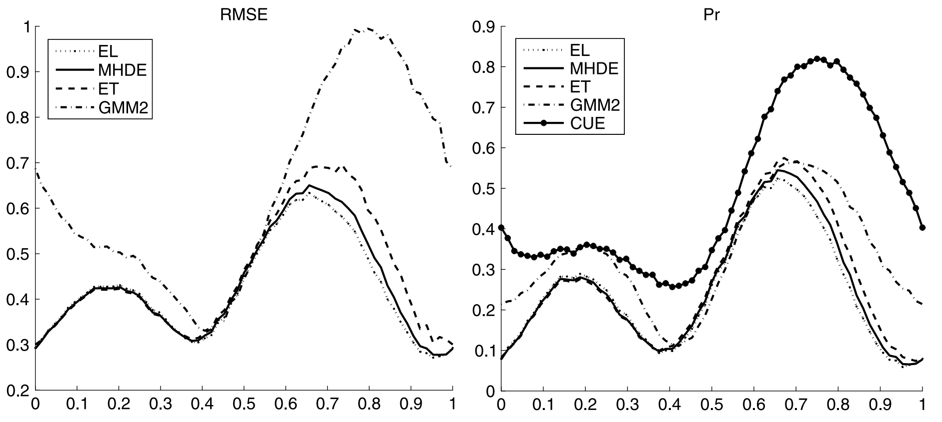

FIGURE 1.—Local neighborhood of the true model. "Pr" denotes $\Pr\{|\hat{\theta}-\theta_0|>0.5\}$.

the values of ω's between 0.4 and 0.6, where the performances of the estimators other than CUE are rather close. ET seems to perform a little worse than MHDE and EL.

One needs to be cautious in drawing conclusions based on limited simulation experiments as presented here. Nevertheless, it appears that two general features emerge from our results. First, the GMM type estimators (two step GMM and CUE) tend to be highly sensitive to data perturbations. Applying Beran's (1977) logic that connects the robustness of estimators to the forms of their objective functions, this may be attributed to the fact that the GMM objective function is quadratic and therefore tends to react sensitively to the added noises. Second, EL, MHDE, and ET are relatively well behaved, and their rankings, not surprisingly, vary depending on the simulation design. The performance of MHDE, however, seems more stable compared with that of EL or ET: EL and ET exhibit more instability throughout the different perturbation designs. Note that EL, MHDE, ET, and CUE correspond to the GEL estimator with $\gamma=-1,-\frac{1}{2},0,1$ in equation (2.6) of Newey and Smith (2004). Given the good theoretical robustness property of the MHDE, and the proximity of EL and ET in terms of their γ values, it is interesting to observe the reasonably robust behavior of EL and ET. Note that CUE, whose behavior is quite different from that of the MHDE and thus highly nonrobust, has $\gamma=1$, a value that is much higher than the optimally robust $\gamma=-1/2$ of the MHDE.

5. CONCLUSION

In this paper, we have explored the issue of robust estimation in a moment restriction model. The model is semiparametric and distribution-free, and therefore imposes mild assumptions. Yet it is reasonable to expect that the probability law of observations may have some deviations from the ideal distribution as modeled by the moment restriction model. It is then sensible to seek estimation procedures that are robust against slight perturbations in the probability measure that generates observations, which can be caused by, for example, data contamination. Our main theoretical result shows that the minimum Hellinger distance estimator (MHDE) possesses optimal minimax robust properties. Moreover, it remains semiparametrically efficient when the model assumptions hold. Convenient numerical algorithms for its implementation are provided. Our simulation results indicate that GMM can be highly sensitive to data perturbations. The performance of the MHDE remains stable over a wide range of simulation designs, which is in accordance with our theoretical findings.

The results obtained in this paper are concerned with estimation, though it might be potentially possible to extend our robustness theory to parameter testing problems. It is of practical importance to consider robust methods for testing and confidence interval calculations so that the results of statistical inference for moment restriction models are reliable and not too sensitive to departures from model assumptions. Interestingly, there exists a literature on parametric robust inference based on the MHDE method. We plan to investigate robust testing procedure in moment condition models in our future research.

REFERENCES

BERAN, R. (1977): "Minimum Hellinger Distance Estimates for Parametric Models," *The Annals of Statistics*, 5, 445–463. [1186,1190,1196,1199]

——— (1984): "Minimum Distance Procedures," in *Handbook of Statistics*, ed. by P. Krishnaiah and P. Sen. Amsterdam: Elsevier Science, 741–754. [1196]

BICKEL, P. J. (1981): "Quelques aspects de la statistique robuste", in *Ecole d'Eté de Probabilités de Saint Flour IX 1979*, ed. by P. Hennequin. Berlin: Springer, 1–72. [1195,1196]

BICKEL, P., C. KLASSEN, Y. RITOV, AND J. WELLNER (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins Press. [1192,1193]

DONOHO, D., AND R. LIU (1988): "The "Automatic" Robustness of Minimum Distance Functionals," *The Annals of Statistics*, 16, 552–586. [1196]

HALL, P., AND J. L. HOROWITZ (1996): "Bootstrap Critical Values for Tests Based on Generalized Method of Moments Estimators," *Econometrica*, 64, 891–916. [1198]

HANSEN, L. P. (1982): "Large Sample Properties of Generalized Methods of Moments Estimators," *Econometrica*, 50, 1029–1054. [1187]

IMBENS, G. W., R. H. SPADY, AND P. JOHNSON (1998): "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333–357. [1191]

KITAMURA, Y. (1998): "Comparing Misspecified Dynamic Econometric Models Using Nonparametric Likelihood," Working Paper, Department of Economics, University of Wisconsin. [1195]

——— (2002): "A Likelihood-Based Approach to the Analysis of a Class of Nested and Non-Nested Models," Working Paper, Department of Economics, University of Pennsylvania. [1195]

——— (2006): "Empirical Likelihood Methods in Econometrics: Theory and Practice," in *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, ed. by R. Blundell, W. K. Newey, and T. Persson. Cambridge: Cambridge University Press. [1190]

KITAMURA, Y., AND M. STUTZER (1997): "An Information Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65 (4), 861–874. [1191]

KITAMURA, Y., T. OTSU, AND K. EVDOKIMOV (2013): "Supplement to 'Robustness, Infinitesimal Neighborhoods, and Moment Restrictions'," *Econometrica Supplemental Material*, 81, http://www.econometricsociety.org/ecta/Supmat/8617_proofs.pdf; http://www.econometricsociety.org/ecta/Supmat/8617_data_and_programs.zip. [1188]

LECAM, L., AND G. YANG (1990): *Asymptotics in Statistics: Some Basic Concepts*. New York: Springer. [1195]

NEWEY, W. K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135. [1192]

NEWEY, W. K., AND R. J. SMITH (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–255. [1191,1196,1199]

POLLARD, D. (2002): *A User's Guide to Measure Theoretic Probability*. Cambridge: Cambridge University Press. [1190]

REISS, R.-D. (1989): *Approximate Distributions of Order Statistics*. New York: Springer-Verlag. [1190]

RIEDER, H. (1994): *Robust Asymptotic Statistics*. New York: Springer-Verlag. [1186,1192,1196, 1197]

SCHENNACH, S. M. (2007): "Point Estimation With Exponentially Tilted Empirical Likelihood," *The Annals of Statistics*, 35, 634–672. [1195,1196]

SMITH, R. J. (1997): "Alternative Semi-Parametric Likelihood Approaches to Generalized Method of Moments Estimation," *Economic Journal*, 107, 503–519. [1191]

WHITE, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25. [1195]

ZHANG, T. (2006): "From ε-Entropy to KL-Entropy: Analysis of Minimum Information Complexity Density Estimation," *The Annals of Statistics*, 34, 2180–2210. [1190]

Cowles Foundation for Research in Economics, Yale University, New Haven, CT 06520, U.S.A.; yuichi.kitamura@yale.edu,

Dept. of Economics, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, U.K.; t.otsu@lse.ac.uk,

and

Dept. of Economics, Princeton University, Princeton, NJ 08544, U.S.A.; kevdokim@princeton.edu.

Manuscript received June, 2009; final revision received August, 2012.