SHARP BOUNDS UNDER CONTAMINATED OR CORRUPTED SAMPLING WITH VERIFICATION, WITH AN APPLICATION TO ENVIRONMENTAL POLLUTANT DATA

Jeff Dominitz and Robert P. Sherman[1]

July, 2003

Abstract

Let F denote a distribution of interest and G a possibly spurious distribution. This paper derives and nonparametrically estimates sharp bounds on characteristics of F when the data are a mixture of F and G, and a fraction of the data is verified to be from F. Contaminated and corrupted mixtures, with and without monotonicity restrictions, are analyzed. The methods are particularly useful in analyzing environmental pollutant measurements obtained using gas chromatography-mass spectroscopy. Results are applied to measurements of organic pollutant concentrations from the Love Canal. We argue that a corruption with monotonic verification model may be the most appropriate model for this type of data.

Key Words
Gas chromatography-mass spectroscopy; mixture models.

Jeff Dominitz is an assistant professor of economics and public policy in the H. John Heinz III School of Public Policy and Management at Carnegie Mellon University, Pittsburg, PA 15213-3890 (dominitz@andrew.cmu.edu). Robert P. Sherman (sherman@amdg.caltech.edu) is an associate professor of economics and statistics in the Division of Humanities and Social Sciences at the California Institute of Technology, Pasadena, CA 91125 (sherman@amdg.caltech.edu).

1. Introduction

[1]We thank Diane Lambert for providing the organic pollutant data analyzed in this paper. (This data is available from us upon request.) We also thank Nathan Dalleska, Director of the Environmental Analysis Center at the California Institute of Technology, for sharing his insights about gas chromatography-mass spectroscopy.

It is often reasonable, in practice, to assume that data are generated according to a mixture model, where each observation is selected from either a distribution of interest, say F, or another, possibly spurious distribution, say G. If selection is independent of the value drawn from F, the data are said to be contaminated. Otherwise, the data are said to be corrupted, a more serious deficiency. In either case, unless untestable assumptions about the data generating process hold, it is not possible to identify characteristics of F such as moments, probabilities, and quantiles. However, given a lower bound on the probability of selection from F, Horowitz and Manski (1995) identify sharp bounds on such characteristics under either contaminated or corrupted sampling. (Bounds are sharp if they are attainable under the maintained assumptions about the data generating mechanism.) They also show how to nonparametrically estimate these bounds.

Sometimes there is more information than simply a lower bound on the probability of selection from F. Sometimes a subset of the sample is verified to be drawn from F. We say that data generated in this way come from a mixture model with verification.

Mixture models with verification can be used to model many interesting data problems. See Dominitz and Sherman (2003) for an application to school test score data where a significant proportion of students are classified as having limited English language skills. A discussion is also given of applications to data on self-reported income partially supplemented with administrative records, missing data problems, and treatment effects with partially missing treatments. In each of these examples, it is reasonable to assume that a fraction of the data is drawn from a distribution of interest, while the rest of the data may or may not be drawn from this distribution.

The application guiding the methodological development in this paper concerns the analysis of measurements of organic pollutant concentrations using gas chromatography and mass spectroscopy. This application was previously considered in a paper by Lambert and Tierney (1997),

who analyzed measurements obtained from the Love Canal in upstate New York in 1988. The Love Canal was the site of a major environmental disaster.

As discussed in Lambert and Tierney (1997), a gas chromatograph (GC) is supposed to isolate a pollutant of interest from an environmental sample containing many chemical compounds. Since the GC sometimes isolates the wrong pollutant, a mass spectrometer (MS) is used to test whether an isolated compound is the right pollutant. The MS also measures the pollutant concentration. The distribution of interest, F, is the distribution of concentrations from the right pollutant. Problems arise because verification is partial and may depend on concentration level. Chemists believe that verified concentrations are from F, but unverified concentrations may or may not be from F. Moreover, it is believed that lower concentrations from F are likely to be harder to verify than higher concentrations. Thus, there may be information on F in the unverified measurements, while the verified measurements may not be a random sample from F.

Lambert and Tierney (1997) derive sharp bounds on a general functional of F under a contamination with verification model (they do not use this terminology). They propose a nonparametric maximum likelihood estimation procedure and apply this procedure to estimate bounds on mean concentrations of organic pollutants extracted from the Love Canal. Their procedure involves nonparametric estimation of the conditional probability of verification given the level of pollution concentration. Because of this, the procedure has difficulty handling the censoring that occurs in the Love Canal data at very low pollution concentrations. Nor can it naturally incorporate the plausible monotonicity condition mentioned above. The authors call for extensions that could better handle censoring as well as continuous covariates. Their methods do not allow for corruption in the sampling process.

In this paper, we derive and nonparametrically estimate sharp bounds on characteristics of

F under contaminated or corrupted sampling with verification. Our approach to establishing bounds is an extension of the approach of Horowitz and Manski (1995) and is quite different from the approach taken by Lambert and Tierney (1997). This leads to nonparametric estimation procedures that are more direct than the procedure of Lambert and Tierney (1997). Our procedures do not require nonparametric estimation of the conditional probability of verification given the level of pollution concentration. As such, they can easily and naturally handle the censoring in the Love Canal data as well as incorporate plausible monotonicity conditions. Our procedures can also be extended in a straightforward way to handle continuous covariates. Perhaps most importantly, as mentioned above, we allow for corruption in the data generating process. We show that under an additional monotonicity condition that is likely to hold for measurements obtained using gas-chromatography-mass spectroscopy in general and the Love Canal data in particular, the contamination with verification model is invalid. We argue, therefore, that a corruption with verification model incorporating all plausible monotonicity conditions may be a more appropriate model for this type of data. Moreover, we show for the Love Canal data that inferences based on such a corruption model can be substantively different from those based on the corresponding contamination model.

The rest of the paper is organized as follows. In Section 2, we formally define a general mixture model with verification, and present and discuss the assumptions under which we derive sharp bounds on various characteristics of F. In Section 3, we derive sharp bounds on characteristics of F under both corruption and contamination with verification, incorporating censoring and plausible monotonicity restrictions. Section 4 defines sample analogs of the population bounds derived in Section 3 and discusses some of their asymptotic properties. We also discuss some of the benefits derived from convexity and concavity of the functions defining the contamination with verification

bounds, as well as extensions to handle continuous regressors. In Section 5, we apply our methods to the Love Canal data and discuss substantive findings. Section 6 summarizes. Proofs are given in an appendix.

2. Mixture Models with Verification

In this section, we formally define a mixture model with verification, and state and discuss assumptions that characterize interesting submodels for which we derive sharp bounds on various functionals of the distribution of interest, such as expectations, quantiles, and probabilities. We relate all definitions to the Love Canal example discussed in the introduction and analyzed in detail in Section 5. Recall that GC and MS stand for gas chromatograph and mass spectrometer, respectively.

Let Y_1 denote an outcome of interest with distribution F. In a mixture model, one observes Y, a mixture of Y_1 and another random variable Y_0, where G, the distribution of Y_0, may be quite different from F. Let Z denote a selection indicator. If $Z = 1$ then Y is drawn from F. If $Z = 0$, then Y is drawn from G. That is,

$$Y = Y_1 Z + Y_0(1 - Z)\,. \tag{1}$$

Finally, let V denote an observed verification indicator. If $V = 1$, then Y is verified to be drawn from F. If $V = 0$, then Y may be drawn from either F or G. For the Love Canal data, Y_1 denotes an organic pollution concentration of interest, $Z = 1$ if the GC isolates the pollutant of interest, and $V = 1$ if the MS verification test is passed.

In order to handle response censoring of the form encountered in the Love Canal example, suppose $Y \geq \alpha$ where α is known and $c \geq \alpha$ is a known censoring point. We consider a situation

where Y may or may not be censored if its value lies between α and c. This is the situation governing the Love Canal data, where Y is a chemical concentration measured in parts per billion, $\alpha = 0$, and $c = .2$ parts per billion. Some concentration levels are observed in the range $[0, .2]$, while for others, all that is known is that they fall in this interval. Define

$$\begin{aligned} \underline{Y} &= Y\{Y \text{ is observed}\} + \alpha\{Y \text{ is censored}\} \\ \overline{Y} &= Y\{Y \text{ is observed}\} + c\{Y \text{ is censored}\}\,. \end{aligned}$$

Note that both $\underline{Y}$ and $\overline{Y}$ are observed and the case $c = \alpha$ corresponds to no response censoring. Also, note that these definitions nest the standard case of response censoring below a known censoring point, where Y is always censored if it falls below the censoring point.

In sum, in a mixture model with verification and possible response censoring, each member of the sampling distribution is characterized by a vector $(\underline{Y}, \overline{Y}, V, Z, Y_1, Y_0)$, where $\underline{Y}$, $\overline{Y}$, and V are observed. A vector of covariates may also be observed, but we do not formally consider this possibility in this paper.

Let M denote a known, real-valued function on the support of Y_1. Our objective is to derive sharp bounds on $I\!E M(Y_1)$. By choosing $M(s) = s^k$, we can obtain sharp bounds on arbitrary moments of F. By choosing $M(s) = \{s \leq t\}$ for $t \in I\!R$ we can obtain sharp bounds on the cumulative distribution function $F(t)$, which can be inverted to get sharp bounds on arbitrary quantiles of F.

We shall select from the following assumptions to define interesting submodels of the basic mixture model with verification.

A1. $I\!P\{V = 1\} > 0$.

A2. $I\!P\{Z = 1 \mid V = 1\} = 1$.

A3. $I\!E M(Y_1) = I\!E[M(Y_1) \mid Z = 1]$.

A4. $I\!E[M(Y_1) \mid V = 1] \geq I\!E[M(Y_1) \mid V = 0]$.

Assumption A1 says that the probability of verification is positive. By A2, all the verified data are from F. Assumptions A1 and A2 characterize the least restrictive corruption with verification model that we study in this paper. These are reasonable assumptions for the Love Canal data.

Assumption A3 is a mean independence assumption. Assumptions A1, A2, and A3 characterize the least restrictive contamination with verification model that we study here. Notice that A3 follows from the following stronger independence assumption:

B3. Z is independent of Y_1.

Assumption B3 says that selection is independent of the value drawn from F. Lambert and Tierney (1997) assume B3 in establishing their contamination with verification bounds. We get the same sharp bounds whether we assume A3 or B3.

Assumption A4 is a monotonicity restriction that can be placed on either the basic corruption with verification model or the basic contamination with verification model to tighten bounds on functionals of F. Lemma 1A in the appendix shows that if M is increasing, then A4 follows from A1, A2, and the following monotonicity conditions:

B4. $I\!P\{V = 1 \mid Z = 1, Y_1 = y\}$ is increasing in y.

B4′. $I\!P\{Z = 1 \mid Y_1 = y\}$ is increasing in y.

In their analysis of the Love Canal data, Lambert and Tierney (1997) assume a contamination with verification model and impose monotonicity restriction B4, which allows that smaller concentrations of the pollutant of interest are likely to be harder to verify than larger concentrations. In

short, they assume a contamination with verification model characterized by assumptions A1, A2, B3, and B4.

It may also be reasonable to assume that smaller concentrations of the pollutant of interest are likely to be harder to isolate than larger concentrations. (We learned of the plausibility of this assumption through discussions about GC-MS technology with Nathan Dalleska, Director of the Environmental Analysis Center at the California Institute of Technology.) Assumption B4′ formally allows for this. Notice that if B4′ holds nontrivially, then B3 cannot hold, whereas A1, A2, B4, and B4′ imply A4. Thus, a corruption with verification model under assumptions A1, A2, and A4 may be a more appropriate model for the Love Canal data than a contamination with verification model under assumptions A1, A2, B3, and B4. We return to this consideration in Section 5.

3. Sharp Bounds

In this section, we derive sharp bounds on $I\!E M(Y_1)$ for various choices of M and for various mixture models with verification. We allow response censoring of the form defined in Section 2.

We begin with the basic corruption with verification model characterized by assumptions A1 and A2. Let M be a known, real-valued function on $I\!R$, and suppose the support of $M(Y_1)$ is contained in the interval $[a, b]$, where a and b are known and satisfy $-\infty < a < b < \infty$. For the Love Canal data, when $M(s) = s$, $a = 0$ and $b = 10^9$. When $M(s) = \{s \le t\}$, $a = 0$ and $b = 1$. Write v for $I\!P\{V = 1\}$.

Theorem 1. *If A1 and A2 hold, then* $\lambda_1 \le I\!E M(Y_1) \le u_1$ *where*

$$\begin{aligned} \lambda_1 &= I\!E[M(\underline{Y}) \mid V = 1]v + a(1 - v) \\ u_1 &= I\!E[M(\overline{Y}) \mid V = 1]v + b(1 - v)\,. \end{aligned}$$

Moreover, these bounds are sharp.

Next, we consider adding the monotonicity restriction A4 to the basic corruption with verification model. This restriction can have considerable identifying power in applications where there is a very large upper bound on $M(Y_1)$, as is the case with the Love Canal data, where $b = 10^9$ when $M(s) = s$.

THEOREM 2. *If A1, A2, and A4 hold, then* $\lambda_2 \leq I\!E M(Y_1) \leq u_2$ *where*

$$\begin{aligned} \lambda_2 &= \lambda_1 \\ u_2 &= I\!E[M(\overline{Y}) \mid V = 1]\,. \end{aligned}$$

Moreover, these bounds are sharp.

REMARK 1. If the sense of the inequality in A4 is reversed, then it is easy to show that $\lambda_2 = I\!E[M(\underline{Y}) \mid V = 1]$ and $u_2 = u_1$ are sharp bounds on $I\!E M(Y_1)$.

Next, we consider the basic contamination with verification model characterized by assumptions A1, A2, and A3, and assume that M is increasing. For ease of exposition, we also assume that Y is continuously distributed (as it is for the Love Canal data).

By A3, $I\!E M(Y_1) = I\!E[M(Y_1) \mid Z = 1] = I\!E[M(Y) \mid Z = 1]$. It follows from A1 and A2 that $I\!E[M(Y) \mid Z = 1, V = 1] = I\!E[M(Y) \mid V = 1]$. Write p^* for $I\!P\{V = 1 \mid Z = 1\}$. Then

$$I\!E M(Y_1) = I\!E[M(Y) \mid V = 1]p^* + I\!E[M(Y) \mid Z = 1, V = 0](1 - p^*)\,. \tag{2}$$

If we can find a lower bound on $I\!P\{Z = 1 \mid V = 0\}$, then we can apply Proposition 4 in Horowitz

and Manski (1995) to bound $I\!E[M(Y) \mid Z = 1, V = 0]$, which, in turn, will lead to bounds on $I\!EM(Y_1)$ in (2).

To this end, let $\pi(p^*) = I\!P\{Z = 1 \mid V = 0\}$. Recall that $v = I\!P\{V = 1\}$. A1, A2, and Bayes' rule imply that $\pi(p^*) = [(1-p^*)v]/[p^*(1-v)]$. Recall the definitions of $\underline{Y}$ and $\overline{Y}$ given in Section 2. Let U denote a $U[0,1]$ random variable independent of Y. Define

$$\begin{aligned}
\underline{\mathcal{Y}} &= (\alpha + U)\{Y \text{ is censored}\} + (Y+1)\{Y \text{ is observed}\} \\
\overline{\mathcal{Y}} &= Y\{Y \text{ is observed}, Y < c\} + (c+U)\{Y \text{ is censored}\} + (Y+1)\{Y \text{ is observed}, Y > c\}.
\end{aligned}$$

Note that $\underline{\mathcal{Y}}$ and $\overline{\mathcal{Y}}$ are continuous analogues of $\underline{Y}$ and $\overline{Y}$, respectively. Write $\underline{Q}$ for the population quantile function of $\underline{\mathcal{Y}}$ given $V = 0$ and $\overline{Q}$ for the population quantile function of $\overline{\mathcal{Y}}$ given $V = 0$. By Proposition 4 in Horowitz and Manski (1995), the interval

$$\left[I\!E[M(\underline{Y}) \mid \underline{\mathcal{Y}} \leq \underline{Q}(\pi(p^*)), V = 0], I\!E[M(\overline{Y}) \mid \overline{\mathcal{Y}} > \overline{Q}(1 - \pi(p^*)), V = 0]\right] \tag{3}$$

contains $I\!E[M(Y) \mid Z = 1, V = 0]$. Combining (3) with (2) produces bounds on $I\!EM(Y_1)$. However, these bounds are infeasible since p^* is unknown.

To develop feasible bounds, first note that $v \leq p^* \leq 1$. This follows from A1, A2, and Bayes' rule. For each $p \in [v, 1]$, define $\pi(p) = [(1-p)v]/[p(1-v)]$. Define lower and upper bound functions

$$\begin{aligned}
L(p) &= pI\!E[M(\underline{Y}) \mid V = 1] + (1-p)I\!E[M(\underline{Y}) \mid \underline{\mathcal{Y}} \leq \underline{Q}(\pi(p)), V = 0] \\
U(p) &= pI\!E[M(\overline{Y}) \mid V = 1] + (1-p)I\!E[M(\overline{Y}) \mid \overline{\mathcal{Y}} > \overline{Q}(1 - \pi(p)), V = 0].
\end{aligned}$$

This leads to the following result.

THEOREM 3. *If A1, A2, and A3 hold, then* $\lambda_3 \leq I\!E M(Y_1) \leq u_3$ *where*

$$\begin{aligned} \lambda_3 &= \inf_{p \in [v,1]} L(p) \\ u_3 &= \sup_{p \in [v,1]} U(p) \,. \end{aligned}$$

Moreover, these bounds are sharp.

Next, we consider the effect of imposing the monotonicity assumption A4 on the basic contamination model.

THEOREM 4. *If A1, A2, A3, and A4 hold, then* $\lambda_4 \leq I\!E M(Y_1) \leq u_4$ *where*

$$\begin{aligned} \lambda_4 &= \lambda_3 \\ u_4 &= I\!E[M(\overline{Y}) \mid V = 1] \,. \end{aligned}$$

Moreover, these bounds are sharp.

REMARK 2. Note that $u_2 = u_4$. That is, imposing A4 leads to the same upper bound under either corruption with verification or contamination with verification. As before, if the sense of the inequality in A4 is reversed, then $\lambda_4 = I\!E[M(\underline{Y}) \mid V = 1]$ and $u_4 = u_3$ are sharp bounds on $I\!E M(Y_1)$.

4. ESTIMATION

In this section, we develop sample analogs of the population bounds for the mixture models with verification developed in Section 3. We also discuss some of their asymptotic properties, some

of the benefits derived from convexity and concavity of the functions defining the contamination with verification bounds, and extensions covering continuous regressors.

We begin by developing sample analogues of the corruption with verification bounds established in Section 3. Let $(\underline{Y}_i, \overline{Y}_i, V_i, Z_i, Y_{1i}, Y_{0i})$, $i = 1, \ldots, n$, denote iid draws from the mixture model with verification defined in Section 2. Define $n_1 = \sum_{i=1}^{n} V_i$ and $\hat{v} = n_1/n$.

Refer to Theorem 1 in Section 3. Define

$$\begin{aligned} \hat{\lambda}_1 &= \hat{v} \sum_{i=1}^{n} M(\underline{Y}_i) V_i / n_1 + a(1 - \hat{v}) \\ \hat{u}_1 &= \hat{v} \sum_{i=1}^{n} M(\overline{Y}_i) V_i / n_1 + b(1 - \hat{v}) \,. \end{aligned}$$

Next, refer to Theorem 2 in Section 3. Define

$$\begin{aligned} \hat{\lambda}_2 &= \hat{\lambda}_1 \\ \hat{u}_2 &= \sum_{i=1}^{n} M(\overline{Y}_i) V_i / n_1 \,. \end{aligned}$$

A straightforward application of the delta method shows that all the estimators defined above are $\sqrt{n}$-consistent estimators of their population counterparts and asymptotically normally distributed.

Next, we develop sample analogues of the contamination with verification bounds established in Section 3. Define $n_0 = n - n_1$ and $\hat{\pi}(p) = [(1-p)\hat{v}]/[p(1-\hat{v})]$. Recall the definitions of $\underline{\mathcal{Y}}$ and $\overline{\mathcal{Y}}$ given in Section 3. Let $\underline{\mathcal{Y}}_i$ and $\overline{\mathcal{Y}}_i$, $i = 1, \ldots, n$, denote the corresponding sample quantities. Write $\hat{\underline{Q}}$ for the empirical quantile function of the $\underline{\mathcal{Y}}_i$'s for V_i's equal to zero and $\hat{\overline{Q}}$ for the empirical quantile function of the $\overline{\mathcal{Y}}_i$'s for V_i's equal to zero.

Refer to Theorem 3 in Section 3. Define sample lower and upper bound functions

$$\begin{aligned}
\hat{L}(p) &= p\sum_{i=1}^{n} M(\underline{Y}_i)V_i/n_1 + (1-p)\sum_{i=1}^{n} M(\underline{Y}_i)(1-V_i)\{\underline{\mathcal{Y}}_i \le \hat{\underline{Q}}(\hat{\pi}(p))\}/\hat{\pi}(p)n_0 \\
\hat{U}(p) &= p\sum_{i=1}^{n} M(\overline{Y}_i)V_i/n_1 + (1-p)\sum_{i=1}^{n} M(\overline{Y}_i)(1-V_i)\{\overline{\mathcal{Y}}_i > \hat{\overline{Q}}(1-\hat{\pi}(p))\}/\hat{\pi}(p)n_0\,.
\end{aligned}$$

Define extreme value estimators

$$\begin{aligned}
\hat{\lambda}_3 &= \inf_{p\in[\hat{v},1]} \hat{L}(p) \\
\hat{u}_3 &= \sup_{p\in[\hat{v},1]} \hat{U}(p)\,.
\end{aligned}$$

Next, refer to Theorem 4 in Section 3. Define

$$\begin{aligned}
\hat{\lambda}_4 &= \hat{\lambda}_3 \\
\hat{u}_4 &= \sum_{i=1}^{n} M(\overline{Y}_i)V_i/n_1\,.
\end{aligned}$$

Arguments in Sherman (2003) show that if Y is bounded (as it is in the Love Canal example), then $\hat{\lambda}_3$ and $\hat{u}_3$ are $\sqrt{n}$-consistent estimators of their population counterparts. The presence of the empirical quantile functions in the indicator function factors of the summands comprising $\hat{L}(p)$ and $\hat{U}(p)$ makes establishing the limiting distribution of $\hat{\lambda}_3$ and $\hat{u}_3$ a difficult task. Dominitz and Sherman (2003) show that estimators analogous to $\hat{\lambda}_3$ and $\hat{u}_3$ are asymptotically normally distributed when Y has a finite number of support points. Their arguments do not apply to estimators based on the Love Canal data, where $\underline{Y}$ and $\overline{Y}$ are mixtures of a continuous response and a point mass at a censoring threshold. We leave for future work the establishment of the limiting distribution of these estimators when the response is continuously distributed or a mixture

of a continuous and discrete distribution.

Arguments in Dominitz and Sherman (2003) can be adapted to show that $\hat{L}(p)$ is a piecewise linear convex function and $\hat{U}(p)$ is a piecewise linear concave function on $[\hat{v}, 1]$. Moreover, the ordinates of the n possible kink points of $\hat{L}(p)$ are given by $\hat{\pi}^{-1}(\underline{\hat{H}}(y_i))$, $i = 1, \ldots, n$, where $\hat{\pi}^{-1}(p) = \hat{v}/[\hat{v}+p(1-\hat{v})]$, $\underline{\hat{H}}$ is the empirical distribution function of the $\underline{Y}_i$'s, and y_i, $i = 1, \ldots, n$ are the realized values of the $\underline{Y}_i$'s. Similarly, the ordinates of the n possible kink points of $\hat{U}(p)$ are given by $\hat{\pi}^{-1}(1 - \overline{\hat{H}}(y_i))$, $i = 1, \ldots, n$, where $\overline{\hat{H}}$ is the empirical distribution function of the $\overline{Y}_i$'s, and y_i, $i = 1, \ldots, n$ are the realized values of the $\overline{Y}_i$'s. Thus, to find $\hat{\lambda}_3$ or u_3, it is sufficient to evaluate the corresponding criterion functions at the corresponding ordinates, which are computable. However, since $\hat{L}(p)$ is convex and $\hat{U}(p)$ is concave, a binary search over the corresponding ordinates will find $\hat{\lambda}_3$ and u_3 after only $O(\log n)$ function evaluations. This can result in substantial savings in computation time when (i) n is large, (ii) bootstrap estimates of the distribution of the extreme value estimators are desired, or (iii) the response depends on a vector of regressors, and it is desirable to evaluate the extreme value estimators for many regressor values (see below).

Finally, we note that the estimation procedures developed in this section apply immediately to the situation where the response variable Y depends on a vector of discrete regressors, X. One simply restricts attention to the subset of the data corresponding to a possible value of X and proceeds as prescribed in this section. This is illustrated in the Love Canal example analyzed in Section 5, where 6 different labs analyze randomly drawn soil samples from the Love Canal and comparison regions. In this example, X denotes lab and can take on one of six possible values.

The extension to the situation where X has continuous components is straightforward. One simply replaces all sample averages defined in this section with standard nonparametric regression analogs. Sherman (2003) shows how and develops primitive conditions implying convergence of

the extreme value estimators at rate $n^{-2/[d+4]}$, where d is the number of continuous components of X. Conditions implying asymptotic normality of the corruption with verification extreme value estimators are also developed.

5. An Example: The Love Canal Study

In this section, we illustrate the usefulness of our results on corruption and contamination with verification by reanalyzing data on environmental pollutants 2-chloronapthalene (2-CNAP) and α-BHC in both the Love Canal and a comparison region. This data was previously analyzed by Lambert and Tierney (1997).

As described by Lambert and Tierney, the measurement and verification process for each compound proceeded as follows. Soil samples were collected from a number of randomly-selected locations in the Love Canal and in a comparison region. Samples were then randomly assigned to labs for pollutant analysis by gas chromatography-mass spectroscopy. The gas chromatograph was used to attempt to isolate the pollutant of interest. The mass spectrometer was used to ionize the isolated compound and then (1) verify that the compound was the pollutant of interest and (2) measure the concentration of the most abundant ion. The verification procedure yielded reports of either verified ($V = 1$) or not verified ($V = 0$). At each lab, a positive fraction was verified, and so assumption A1 holds.

According to Lambert and Tierney, "Chemists believe that all of the verified GC-MS measurements belong to the pollutant of interest, but unverified GC-MS measurements may or may not belong to the pollutant" (page 937). If this statement is correct, then A2 holds. Lambert and Tierney further assume B3, that isolation of the compound is independent of its concentration. They also assert the plausibility of assumption B4, that the probability of verification when the correct pollutant has been isolated is increasing in the concentration level. Assumptions A1, A2,

and B3 characterize their contamination with verification model. Assumptions A1, A2, B3, and B4 characterize their contamination with monotonic verification model. (In this paper, we characterize contamination with verification with the weaker set A1, A2, and A3, and contamination with monotonic verification with the set A1, A2, A3, and A4.)

As discussed at the end of Section 2, it may also be reasonable to assume B4′, that the probability of isolation is increasing in the concentration level. If B4′ holds nontrivially, then B3 cannot hold. This calls into question a model based on B3 and suggests that a corruption with monotonic verification model characterized by assumptions A1, A2, and A4 (recall from Section 2 that A4 follows from A1, A2, B4, and B4′) may be more appropriate for this data. Moreover, in this section, we show that corruption with monotonic verification bounds can lead to inferences about pollution concentrations that are substantively different from inferences based on contamination with monotonic verification bounds. This underscores the need for corruption with verification bounds.

We also note that, unlike the estimator proposed in Section 4 for the contamination with verification model, the nonparametric maximum likelihood estimator derived by Lambert and Tierney requires kernel smoothing even in the absence of continuous covariates. This approach does not allow them to easily handle the response censoring that is present in the Love Canal data or to add continuous covariates to the analysis. In addition, Lambert and Tierney can only incorporate assumption B4 in estimation with an "ad hoc" adjustment requiring that an "unverified measurement is tied with [equal to] the smallest verified measurement" (p. 940). We shall illustrate the identifying power of assumptions A3 and A4 and the ease with which these assumptions and censoring can be incorporated into our estimation procedure.

Measurements were reported in parts per billion (ppb). However, when unverified measurements

were below 0.2 ppb, the protocol allowed the lab to simply report "less than .2 ppb." Unlike the approach taken by Lambert and Tierney, our bounds easily incorporate such censoring. Table 1 presents descriptive statistics for the data produced by each of six labs, as well as for the aggregated data. Note that bounds on the sample means of measured concentrations are reported when some observed values are censored. These bounds are obtained by assigning to each censored observation the lower limit value 0.00 to calculate the lower bound and the upper limit value 0.20 to calculate the upper bound.

Several other features of the table are noteworthy. The proportion of verified measurements $n_1/n = \hat{v}$ varies greatly across labs, ranging from 0.32 to 0.97 for 2-CNAP and from 0.29 to 0.87 for α-BHC. The proportion censored among the unverified also varies greatly, ranging 0.00 to 0.50 for 2-CNAP and from 0.00 to 0.86 for α-BHC. Note that not all unverified values below 0.2 are censored. For α-BHC, the verified measurements tend to be much larger than the unverified measurements, but this relationship is not clear for 2-CNAP, which has mean verified concentration levels below the censoring value $c = 0.2$ at each lab.

Consider now estimates of the distribution of 2-CNAP concentration in Love Canal using data for all labs combined. Let $M(Y_1) = \{Y_1 \leq t\}$ for $t \in R$. Figure 1 presents our estimates of the corruption and contamination bounds with monotonic verification for a fine grid of values of t on the interval $[0, 0.3]$; that is, estimated bounds on the cumulative distribution function (cdf) over this interval. Note that since $M(Y_1) = \{Y_1 \leq t\}$ is monotone decreasing, the direction of the inequality in A4 is reversed, and so this assumption affects the lower bounds rather than the upper bounds. That is, $\hat{\lambda}_1 \leq \hat{\lambda}_2$ and $\hat{\lambda}_3 \leq \hat{\lambda}_4$, but $\hat{u}_1 = \hat{u}_2$ and $\hat{u}_3 = \hat{u}_4$ (see Remark 1 and Remark 2 in Section 3).

We see that the lower bounds coincide (i.e., $\hat{\lambda}_2 = \hat{\lambda}_4$) and the corruption upper bounds $\hat{u}_2$

exceed the contamination upper bounds $\hat{u}_4$. For example, the bounds on $P\{Y_1 \leq 0.070\}$ are $[0.44, 0.60]$ under corruption versus $[0.44, 0.54]$ under contamination. Note also that the lower bound on $P\{Y_1 \leq 0\}$ equals 0.00 under either corruption or contamination, because all measurements that may equal 0 are unverified censored values, so it may be that there are no true zero concentrations of 2-CNAP. In contrast, the corruption upper bound at $t = 0$ is $\hat{u}_2 = 1 - \hat{v} = 0.28$. The contamination upper bound at $t = 0$ is just $\hat{u}_4 = 0.02$. The vertical distance between the estimated upper bounds on the cdf is an indicator of the identifying power of assumption A3. This distance is maximized at $t = 0$ and decreases monotonically with t.

Note also that we may invert the bounds on the cdf to obtain bounds on quantiles. For instance, the bounds on the median of Y_1 are $[0.055, 0.078]$ under corruption versus $[0.066, 0.078]$ under contamination. Note that the endpoints of these intervals are the ordinates of the points of intersection of the estimated cdfs and the horizontal line through 0.50 in Figure 1. We also report these values in the upper-right panel of Table 2, along with estimates of joint 90%-confidence intervals based on 100 bootstrap estimates $(\hat{\lambda}^*, \hat{u}^*)$. Following Horowitz and Manski (2000), the estimated intervals are obtained by finding the smallest value z^* such that $\Pr(\hat{\lambda}^* - z^* \leq \hat{\lambda}, \hat{u} \leq \hat{u}^*) = 0.90$. The estimated confidence intervals for bounds on the population median of Y_1 are $[0.049, 0.084]$ and $[0.061, 0.084]$ under corruption with monotonic verification and contamination with monotonic verification, respectively. Thus, maintaining assumption A3 reduces the width of the point estimate of the bounds on the median by almost one-half and the width of the confidence interval by over one-third.

Figure 2a illustrates the calculation of contamination bounds on the mean concentration of 2-CNAP in Love Canal, based on Lab 2 measurements. Figure 2b uses Lab 6 measurements. The pictures appear different because Lab 2 has no censored observations, whereas 6.5 percent of the

Lab 6 unverified observations are censored. Note the convexity and concavity of the respective lower and upper bound functions in both figures.

In Figure 2a, $\hat{U}(\hat{v}) = \hat{L}(\hat{v}) = 0.064$, which is the sample mean of observed measurements. As p increases, $\hat{U}(p)$ increases monotonically until $p = 0.903$, with $\hat{U}(0.903) = \hat{u}_3 = 0.068$. From there, $\hat{U}(p)$ decreases monotonically until $\hat{U}(1) = \hat{L}(1) = 0.065$, which is the sample mean of verified measurements $\hat{E}[Y|V = 1]$. Similarly, $\hat{L}(p)$ decreases monotonically from either direction, reaching its infimum at $p = 0.871$, with $\hat{L}(0.871) = \hat{\lambda}_3 = 0.062$. Thus, we have $0.062 \leq \hat{E}[Y_1] \leq 0.068$.

Now, suppose we invoke A4. Then, as depicted in Figure 2a, the upper bound is reduced to $\hat{u}_4 = \hat{E}[Y|V = 1] = 0.065$. This assumption has no effect on the lower bound. Thus, the monotonicity assumption reduces the width of the estimated contamination bounds by about two-fifths. Figure 2b presents similar bounds for Lab 6, where the main qualitative difference arises from the censoring that separates $\hat{L}(\hat{v})$ and $\hat{U}(\hat{v})$. In this case, maintaining A4 reduces the width of the contamination bounds by almost two-thirds.

The rectangles in Figure 3a depict point estimates of the contamination with monotonic verification bounds on the mean concentration of 2-CNAP for each lab for both Love Canal (LC) and the comparison region (CR). Figure 3b presents the corresponding bounds for mean concentration of α-BHC. For example, the bounds $[\hat{\lambda}_4, \hat{u}_4] = [0.062, 0.065]$ in Figure 2a are depicted as the third rectangle from the left (Lab 2, LC) in Figure 3a. The bounds $[\hat{\lambda}_4, \hat{u}_4] = [0.044, 0.049]$ from Figure 2b are reported in the column denoted (Lab 6, LC).

The vertical lines in Figure 3a depict point estimates of the bounds on mean 2-CNAP concentration under corruption with monotonic verification. The identifying power of assumption A3 appears quite striking, as the lines tend to extend relatively far below the rectangle. Of course, the comparison would be even more striking were we to not maintain A4, in which case the upper

bounds on the mean under corruption would approach one billion ppb.

The picture is quite different with respect to mean concentration of α-BHC concentration, described in Figure 3b. Here, the distance between the estimated corruption and contamination lower bounds is almost imperceptible. This occurs because a sizeable fraction of verified observations at each lab are an order of magnitude greater than the median verified observation. In fact, as reported in the bottom panel of Table 2 and depicted in the far right of Figure 3b (All, LC), the estimated bounds on mean concentration in Love Canal under corruption with monotonic verification are $[2.695, 3.549]$, whereas the corresponding bounds on the median are just $[0.1820, 0.2824]$, based on all laboratory measurements. The bounds under contamination (reported in Table 2 and depicted in the Figure 3b) are very similar.

The estimated bounds on the mean α-BHC concentration may yield different inferences about pollution in Love Canal than do the bounds on the median. For instance, suppose one focuses on comparisons across regions of the mean concentration under contamination, as do Lambert and Tierney. The point estimates of the bounds on the mean in Love Canal are $[2.727, 3.549]$, which greatly exceed the bounds for the comparison region of $[0.544, 1.044]$. The estimated 90%-confidence intervals nearly overlap, but not quite–$[1.555, 4.721]$ versus $[0.035, 1.551]$. Further, A4 has no identifying power in this case, as the estimated bounds under contamination are identical with and without the monotonic verification assumption. Under corruption with monotonic verification, the point estimate of the bounds in Love Canal also greatly exceed the estimate in the comparison region, but the confidence intervals overlap slightly.

Now consider a comparison of bounds on the median under contamination. We have a point estimate of $[0.198, 0.297]$ in Love Canal, versus $[0.077, 0.211]$ in the comparison region. The estimated bounds tighten somewhat when A4 is invoked, but they still overlap across regions. Therefore, the

bounds under corruption with monotonic verification overlap as well.

As shown for each lab in Figure 3a and reported overall in Table 2, the estimated bounds on the mean concentration of 2-CNAP in Love Canal under corruption with monotonic verification overlap the comparison region bounds in each laboratory except 1 and 7. Under contamination with monotonic verification, the estimated bounds on the mean in Love Canal exceed the comparison region bounds based on data from Labs 1,6, and 8, and actually fall below the comparison region bounds in Labs 2 and 7. Thus, the contamination and corruption models yield very different findings if one focuses on the point estimates of the bounds.

6. Summary

This paper undertakes a nonparametric analysis of models of corruption with verification and contamination with verification. It extends and improves upon some of the work begun by Lambert and Tierney (1997) on verification problems.

Recall that F denotes the unknown distribution of interest. Lambert and Tierney (1997) derive sharp bounds on functionals of F for a contamination with verification model, and develop nonparametric maximum likelihood-type estimators of these bounds. They do not consider the corruption with verification model. Their estimation procedure does not easily handle censored responses or incorporate plausible monotonicity restrictions. Nor can their procedure be easily extended to handle the case where the response variable depends on continuous regressors.

This paper develops sharp bounds on functionals of F for both corruption with verification and contamination with verification models. Sharp bounds under plausible monotonicity conditions are also derived. Nonparametric estimators of the bounds are developed that easily incorporate censoring and plausible monotonicity restrictions. All estimators are $\sqrt{n}$-consistent, and the estimators

for the corruption with verification model are asymptotically normal. The estimation procedure can be easily extended to handle continuous regressors.

We apply the estimation procedures to organic pollutant data from the Love Canal. These data are partially censored and can be reasonably assumed to satisfy certain monotonicity restrictions. If these restrictions hold, then the contamination with verification model is invalid. Thus, we argue that a corruption with verification model that incorporates censoring and all plausible monotonicity restrictions may be a more appropriate model for this data.

APPENDIX

LEMMA 1A. *Suppose M is an increasing function on the support of Y_1. If A1, A2, B4, and B4′ hold, then A4 holds.*

PROOF. Assumptions A1 and A2 imply that

$$I\!P\{V = 1 \mid Y_1 = y\} = I\!P\{V = 1 \mid Z = 1, Y_1 = y\} I\!P\{Z = 1 \mid Y_1 = y\} .$$

Assumptions B4 and B4′ imply that $I\!P\{V = 1 \mid Y_1 = y\}$ is increasing in y. We will show that this last monotonicity condition implies that for each t in the support of Y_1,

$$I\!P\{Y_1 \leq t \mid V = 1\} \leq I\!P\{Y_1 \leq t\} \leq I\!P\{Y_1 \leq t \mid V = 0\} . \tag{4}$$

The result will follow from this stochastic dominance condition and the fact that M is increasing.

By Bayes' rule,

$$I\!P\{Y_1 \leq t \mid V = 1\} = I\!P\{V = 1 \mid Y_1 \leq t\} I\!P\{Y_1 \leq t\} / I\!P\{V = 1\} .$$

Thus, $I\!P\{Y_1 \le t \mid V = 1\} \le I\!P\{Y_1 \le t\}$ if and only if $I\!P\{V = 1 \mid Y_1 \le t\} \le I\!P\{V = 1\}$. But

$$I\!P\{V = 1\} = I\!P\{V = 1 \mid Y_1 \le t\} I\!P\{Y_1 \le t\} + I\!P\{V = 1 \mid Y_1 > t\} I\!P\{Y_1 > t\} .$$

By the monotonicity condition, $I\!P\{V = 1 \mid Y_1 \le t\} \le I\!P\{V = 1 \mid Y_1 > t\}$, implying that $I\!P\{V = 1 \mid Y_1 \le t\} \le I\!P\{V = 1\}$. The proof that $I\!P\{Y_1 \le t\} \le I\!P\{Y_1 \le t \mid V = 0\}$ is similar. This proves (4). □

PROOF OF THEOREM 1. We establish sharpness of the upper bound. The proof for the lower bound is similar. By A1 and A2, $I\!E[M(Y_1) \mid V = 1] = I\!E[M(Y) \mid V = 1]$. It follows that

$$I\!EM(Y_1) = I\!E[M(Y) \mid V = 1]v + I\!E[M(Y_1) \mid V = 0](1 - v) .$$

Suppose that all censored Y values equal c, so that $Y = \overline{Y}$. In addition, suppose that given $V = 0$, the distribution of $M(Y_1)$ puts unit mass at b. These suppositions are consistent with A1, A2, and the assumptions about the support of $M(Y_1)$. Under these suppositions, $I\!EM(Y_1) = u_1$. This proves the result. □

PROOF OF THEOREM 2. Apply A4 to get

$$\begin{aligned} I\!EM(Y_1) &= I\!E[M(Y_1) \mid V = 1]v + I\!E[M(Y_1) \mid V = 0](1 - v) \\ &\le I\!E[M(Y_1) \mid V = 1] . \end{aligned}$$

By A1 and A2, $I\!E[M(Y_1) \mid V = 1] = I\!E[M(Y) \mid V = 1]$. This yields the upper bound. As in the proof of Theorem 1, suppose that all censored Y values equal c, so that $Y = \overline{Y}$. In addition, suppose that given $V = 0$, the distribution of $M(Y_1)$ puts unit mass at $I\!E[M(\overline{Y}) \mid V = 1]$. Then

$I\!E M(Y_1) = u_2$, proving sharpness. □

PROOF OF THEOREM 3. Recall the definition of $L(p)$ and $U(p)$ given prior to the statement of Theorem 3. From (2) and (3) we obtain the infeasible bounds

$$L(p^*) \leq I\!E M(Y_1) \leq U(p^*)\,.$$

It follows from this and the definition of λ_3 and u_3 that

$$\lambda_3 \leq L(p^*) \leq I\!E M(Y_1) \leq U(p^*) \leq u_3\,.$$

That is, λ_3 and u_3 are bounds for $I\!E M(Y_1)$ under A1, A2, and A3. We want to prove that they are sharp bounds under these assumptions.

Assume, in addition to A1, A2, and A3, that p^* is known. Then $L(p^*)$ and $U(p^*)$ are sharp bounds for $I\!E M(Y_1)$ since they are based on the sharp bounds of Horowitz and Manski (1995) for $I\!E[M(Y)|Z = 1, V = 0]$.

Now drop the assumption that p^* is known. Arguments in Dominitz and Sherman (2003) can be adapted to show that (an extended version of) $\hat{L}(p)$ is convex and (an extended version of) $\hat{U}(p)$ is concave on $[0, 1]$. Arguments in Sherman (2003) show that $\hat{L}(p)$ converges pointwise to $L(p)$ and $\hat{U}(p)$ converges pointwise to $U(p)$ on $[0, 1]$. It then follows from the Convexity Lemma in Pollard (1991) that $L(p)$ is convex and $U(p)$ is concave on $[0, 1]$. Deduce that both $L(p)$ and $U(p)$ are continuous on $[v, 1]$. Since any $p \in [v, 1]$ is a feasible value for p^* under assumptions A1, A2, and A3, so are the p values that minimize $L(p)$ and maximize $U(p)$ over $[v, 1]$. This proves that λ_3 and u_3 are sharp bounds on $I\!E M(Y_1)$. □

PROOF OF THEOREM 4. Argue as in the proof of Theorem 2 to get the upper bound. Argue as in the proofs of Theorem 2 and Theorem 3 to establish sharpness. □

REFERENCES

DOMINITZ, J. AND R. P. SHERMAN (2003): "Nonparametric Analysis of Mixture Models with Verification with an Application to Test Score Data," under review, *Econometrica*.

HOROWITZ, J., AND C. MANSKI (1995): "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63, 281–302.

HOROWITZ, J., AND C. MANSKI (2000): "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data", *Journal of the American Statistical Association*, 95, No. 449, 77–84.

LAMBERT, D. AND L. TIERNEY (1997): "Nonparametric Maximum Likelihood Estimation from Samples with Irrelevant Data and Verification Bias," *Journal of the American Statistical Association*, 92, 937–944.

POLLARD, D. (1991): "Asymptotics for Least Absolute Deviations Regression Estimators," *Econometric Theory*, 7, 186–199.

SHERMAN, R. P. (2003): "Some Asymptotic Results for Bounds Estimation," working paper, California Institute of Technology.