Nonparametric Maximum Likelihood Estimation from Samples with Irrelevant Data and Verification Bias

Diane Lambert
and
Luke Tierney

November 19, 1996

Abstract

Suppose some measurements come from a distribution F that is of interest while others come from another, irrelevant distribution G. Some measurements from F are *verified* and known to be from F. The other, *unverified*, measurements may be from F or from G. Not all measurements from F are equally likely to be verified, and no measurement from G is ever verified. This model applies to measurements of low concentrations obtained using gas chromatography/mass spectroscopy, for example, as is shown in this paper.

How well a feature $T(F)$ of F can be estimated when there are unverified data depends on what can be assumed about F, G, and the conditional probability $v(x)$ of verifying a measurement of x from F. If F, G and v are unrestricted, then more than one choice of (F, G, v) gives the same distribution p of the observable x, and thus $T(F)$ cannot be uniquely estimated from data. But, if the set of values of $T(F)$ that correspond to a distribution p of x is small enough, then it is reasonable to try to estimate that set of $T(F)$ from data. This paper shows that the set of possible values of $T(F)$ is a finite interval for some choices of T, such as the mean, and proposes estimators of the interval of possible values. An example using data from a Love Canal study shows that the partially identified set of $T(F)$ can be sufficiently small for estimators of $T(F)$ to be useful.

Keywords Environmental Data, Identifiability, Nondetect.

Diane Lambert is a Distinguished Member of Technical Staff at Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974. Luke Tierney is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455. He was supported by NSF grants DMS-9005858 and DMS-9303557. The authors thank the associate editor and referees for their insightful comments.

1 Background

Our work is motivated by a problem that arises in measuring low concentrations of organic pollutants with gas chromatography – mass spectroscopy (GC–MS). Typically, an environmental sample contains many compounds, most of which are not of interest. The gas chromatograph first isolates what is hoped to be the pollutant of interest from all other compounds in the environmental sample. The mass spectrometer then ionizes the isolated compound, applies tests to the ions to *verify* that the isolated compound is the pollutant of interest, and converts the amount of the most abundant ion into a concentration. The measured concentration and the pass/fail result from the verification test are recorded by the instrumentation. If the verification test is failed, the measurement is often suppressed and only the word *nondetect* is reported, because the GC–MS verification tests are imperfect at low concentrations. Chemists believe that all the verified GC–MS measurements belong to the pollutant of interest, but unverified GC–MS measurements may or may not belong to the pollutant. The policy of reporting only "nondetect" avoids reporting irrelevant measurements that do not pertain to the pollutant. An obvious question is whether some of the unverified measurements have useful information about the distribution of measurements on the pollutant.

GC–MS samples with unverified measurements are nonstandard. The verified data may not be a random sample of measurements on the pollutant, because smaller measurements on the pollutant are usually harder to verify than larger measurements. On the other hand, the verified data are not just censored from below, because there is no threshold that separates measurements that are always verified from those that are never verified. For example, Figure 1 shows the estimated densities of the verified and unverified measurements from three laboratories for two pollutants, α-BHC and 2-chloronapthalene (2-CNAP), of concern in the 1988 Love Canal study. (See CH2M-Hill, 1988, and Lambert, Peterson and Terpenning, 1991, for more information about the study.) For each laboratory and compound, the ranges of the unverified and verified measurements overlap. Thus, the unverified measurements (called nondetects in the Love Canal study) are not left-censored, and there is no "detection limit" that separates the verified and unverified measurements. For α-BHC, the unverified measurements tend to be smaller than the verified measurements, while for 2-CNAP the unverified measurements resemble the verified measurements.

The remainder of this paper considers a statistical model in which some measurements are unverified, some unverified measurements may be irrelevant, and the probability of verifying a measurement of x from the distribution of interest may depend on x. The statistical model is introduced in Section 2. Identifiability issues that arise when the distributions in the model are not restricted to parametric families are explored in Section 3. New estimators of certain properties of the distribution of interest, such as its mean, are proposed in Section 4 and applied to data from the 1988 Love Canal study in Section 5. Conclusions are given in Section 6.

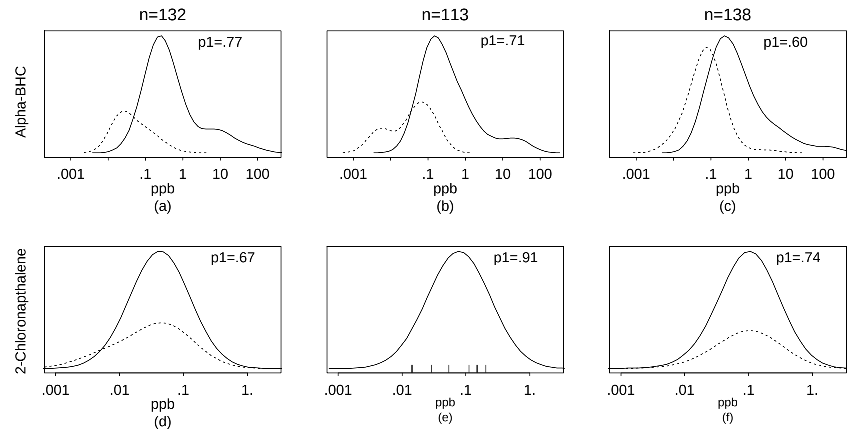


Figure 1: Estimated densities of the verified (solid lines) and unverified (dotted lines) α-BHC and 2-CNAP measurements from three laboratories in the 1988 Love Canal Study. In each plot, n is the total number of measurements and $p1$ the fraction verified. The areas under the verified and unverified curves are proportional to $p1$ and $1 - p1$ respectively. The unverified measurements are shown as points when only a few measurements are unverified.

2 The Statistical Model

There are n independent random vectors (X_i, Y_i, C_i), $i = 1, \ldots, n$ whose joint distribution is defined by

$$\begin{aligned}
C_i &\sim \text{Bernoulli}(\pi), \quad \pi > 0 \\
X_i | C_i &\sim \begin{cases} F & \text{if } C_i = 1 \\ G & \text{if } C_i = 0 \end{cases} \\
Y_i | (X_i = x, C_i) &\sim \begin{cases} \text{Bernoulli}(v(x)) & \text{if } C_i = 1 \\ 0 \text{ with probability one} & \text{if } C_i = 0. \end{cases}
\end{aligned}$$

The average probability $\int v(x)F(dx)$ of verifying a measurement from F is assumed positive. The random variables $C_1, \ldots, C_n$ are not observable. In the context of environmental data, C_i is 1 if the pollutant of interest is isolated, X_i is the measured concentration, and Y_i is 1 if the measurement is verified.

The joint distribution of an observation (X_i, Y_i) is given by

$$\begin{aligned}
p(dx, 1) &= \pi v(x) F(dx) \\
p(dx, 0) &= \pi(1 - v(x))F(dx) + (1 - \pi)G(dx).
\end{aligned}$$

The marginal distribution of X_i is

$$p_X(dx) = p(dx, 1) + p(dx, 0).$$

If $0 < \pi \leq 1$, a verified X_i has distribution

$$p_{X|V}(dx) = \frac{v(x)F(dx)}{\int v(z)\,F(dz)},$$

while if $0 < \pi < 1$ an unverified X_i has distribution

$$p_{X|U}(dx) = \frac{\pi(1 - v(x))F(dx) + (1 - \pi)G(dx)}{1 - \pi \int v(z)\,F(dz)}.$$

When (F, G, v) is restricted to a family of distributions parameterized by θ, maximum likelihood estimates (mle's) of θ and π can be found numerically, *e.g.* by applying the EM algorithm of Dempster, Laird and Rubin (1977). This paper explores estimators of features of F that do not require parametric assumptions on (F, G, v).

3 Identifiability

Is it possible to estimate the mean of F from a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from p when some of the measurements are unverified and nothing is assumed about F, G and v? This section shows that the answer is "not uniquely," no matter how much verified data there is, because many distributions F with different means can generate the same p, and only p can be sampled. On the other hand, p has some information about F, because not all F can generate the same p.

To see the relationship between F and p, first note that

$$p(dx, 1) = \pi v(x)F(dx) \leq \pi F(dx) \leq p_X(dx),$$

which implies that $v(x) = p(dx, 1)/(\pi F(dx))$ satisfies

$$\frac{p(dx, 1)}{p_X(dx)} \leq v(x) \leq 1 \tag{1}$$

for p_X-almost all x. Next, given p, choose any function v that satisfies inequality (1) and take $\tilde{F}(dx) = p(dx, 1)/v(x)$, $\pi = \int \tilde{F}(dx)$, $F(dx) = \tilde{F}(dx)/\pi$, and $G(dx) = (p_X(dx) - \tilde{F}(dx))/(1 - \pi)$. Then F and G are probability distributions and $0 \leq \pi \leq 1$. Thus, given p, each v that satisfies inequality (1) leads to a different F that *generates* p. Given p, let $\mathcal{F}(p)$ be the set of F that can generate p for some distribution G and some verification probability v that satisfies (1).

Although $\mathcal{F}(p)$ contains more than one F, it does not contain all F. If $T(F)$ is a linear functional of F, for example the mean, then its set of values $T(\mathcal{F}(p))$ over $\mathcal{F}(p)$ is a finite interval, as Theorem 1 shows. Its proof is in the appendix.

Theorem 1 *Suppose p is the distribution of the observables (X, Y), and let $\mathcal{F}(p)$ be the set of all distributions F that can generate p. Let $T(F) = \int \phi(x)F(dx)$, where $\phi(x)$ satisfies $\int |\phi(x)|(F(dx)+$*

$G(dx)) < \infty$. Then the set $T(\mathcal{F}(p))$ of values of $T(F)$ that can be obtained from p forms the interval $[T_L, T_U]$, where T_L and T_U are the unique solutions to

$$\begin{aligned} E[\psi_L(\phi(X) - T_L, Y)] &= 0 \\ E[\psi_U(\phi(X) - T_U, Y)] &= 0 \end{aligned}$$

for

$$\psi_L(a, y) = \begin{cases} a & \text{if } y = 1 \text{ or if } y = 0 \text{ and } a \leq 0 \\ 0 & \text{if } y = 0 \text{ and } a > 0 \end{cases}$$

and

$$\psi_U(a, y) = \begin{cases} a & \text{if } y = 1 \text{ or if } y = 0 \text{ and } a \geq 0 \\ 0 & \text{if } y = 0 \text{ and } a < 0. \end{cases}$$

The lower bound T_L is obtained at an F with a density with respect to $p(dx, 1)$ that is proportional to $1/v_L(x)$, where

$$v_L(x) = \begin{cases} p(dx, 1)/p_X(dx) & \text{when } \phi(x) \leq T_L \\ 1 & \text{otherwise.} \end{cases}$$

The upper bound T_U is obtained at an F with a density with respect to $p(dx, 1)$ that is proportional to $1/v_U(x)$, where

$$v_U(x) = \begin{cases} p(dx, 1)/p_X(dx) & \text{when } \phi(x) \geq T_U \\ 1 & \text{otherwise.} \end{cases}$$

The lower bound v_L corresponds to assuming all unverified observations with $\phi(x) \leq T_L$ are from F and all unverified observations with $\phi(x) > T_L$ are from G. The upper bound has a similar interpretation.

Theorem 1 allows the probability of verification $v(x)$ to be arbitrary. Suppose, however, that smaller measurements are harder to verify than larger ones; *i.e.*, that v is increasing. Then the upper bound on $T(F)$ may be smaller than T_U, because $v_U(x)$ is not necessarily monotone. If $p(dx, 1)/p_X(dx)$ does not increase with x, then the lower bound for $T(F)$ may be larger than T_L. Thus, assuming that v is increasing may shorten the interval of possible values of $T(F)$. Theorem 2, which is proved in the appendix, gives the details.

Theorem 2 *Let $\mathcal{F}^*(p)$ be the set of all distributions F that generate p for some distribution G and a nondecreasing v. Suppose ϕ is nondecreasing and $\int |\phi(x)|(F(dx) + G(dx)) < \infty$. Then the set of possible values $T(\mathcal{F}^*(p))$ for $T(F) = \int \phi(x) F(dx)$ forms the interval $[T_L^*, T_U^*]$, where $T_U^* = E[\phi(X)|Y = 1]$ and T_L^* is the unique solution to*

$$E[\psi_L^*(X, T_L^*)Y] = 0 \tag{2}$$

for

$$\psi_L^*(x,c) = \begin{cases} (\phi(x)-c)/q(x) & \textit{when } \phi(x) \le c \\ \phi(x)-c & \textit{otherwise,} \end{cases}$$

where

$$q(x) = \operatorname*{ess\,sup}_{z \le x} \frac{p(dz,1)}{p_X(dz)}$$

and the essential supremum is taken with respect to $p_X(dz)$. The lower bound T_L^* corresponds to

$$v_L^*(x) = \begin{cases} q(x) & \textit{when } \phi(x) \le T_L^* \\ 1 & \textit{otherwise.} \end{cases}$$

The upper bound T_U^* corresponds to $v_U^*(x) = 1$.

To illustrate Theorems 1 and 2, take F to be a normal$(1,1)$ density, G a normal$(0,1)$ density, $\pi = .8$, and $v(x) = exp(x)/(1+exp(x))$. Suppose that only p, rather than (F,G,v,π), is known. Then all that can be inferred about the mean of F is that it lies in the interval $(.716, 1.306)$ (from Theorem 1). No matter how large a random sample is taken from p, we cannot hope to estimate the mean of F precisely but only the interval $(.716, 1.306)$. On the other hand, knowing that the mean of F lies between .716 and 1.306 may be sufficient for some purposes. Assuming that v is nondecreasing implies that the mean of F lies in the slightly shorter interval $(.716, 1.255)$ (Theorem 2). The lower bound does not change, because $p(dx,1)/p_X(dx)$ is increasing.

4 Estimation

Suppose a random sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ is taken from p. Let $t_1 < \cdots < t_K$ be the unique values of $X_1,\ldots,X_n$, and let

$$n_{kj} = \#\{i : (X_i,Y_i) = (t_k,j)\}, \quad n_j = \sum_k n_{kj}, \quad n_{k+} = n_{k0} + n_{k1}, \quad n = n_1 + n_0.$$

As usual, the nonparametric mle of p is the empirical distribution $\hat{p}$, which places mass n_{kj} on (t_k, j) for $k = 1,\ldots,K$ and $j = 0,1$. The mle of the set $\mathcal{F}(p)$ of distributions F consistent with p when no restrictions are placed on v is the set $\mathcal{F}(\hat{p})$ of distributions $\hat{F}$ that are consistent with $\hat{p}$. These distributions $\hat{F}$ are concentrated on $\{t_1,\ldots,t_K\}$ with probabilities $\hat{f}_k = \hat{F}(\{t_k\})$ that satisfy

$$\begin{aligned} \hat{\pi}\hat{v}_k\hat{f}_k &= \hat{p}(\{t_k\},1) = n_{k1}/n \\ \hat{\pi}(1-\hat{v}_k)\hat{f}_k + (1-\hat{\pi})\hat{g}_k &= \hat{p}(\{t_k\},0) = n_{k0}/n \end{aligned}$$

for some choice of $(\hat{G}, \hat{v}, \hat{\pi})$, where $\hat{v}_k = \hat{v}(t_k)$ and $\hat{g}_k = \hat{G}(\{t_k\})$.

Inequality (1) implies that

$$n_{k1}/n_{k+} \le \hat{v}_k \le 1 \tag{3}$$

and that $\hat{f}_k$ is proportional to $\hat{p}(\{t_k\}, 1)/\hat{v}_k$. Thus, the mle $\mathcal{F}(\hat{p})$ of $\mathcal{F}(p)$ can also be expressed as the set of distributions $\hat{F}$ that are concentrated on $\{t_1, \ldots, t_K\}$ and assign probabilities

$$\hat{f}_k = \frac{n_{k1} + a_k n_{k0}}{n_1 + \sum_m a_m n_{m0}}$$

to t_k for $0 \leq a_k \leq 1$. In words, each measurement that is known to be from F is given weight 1 and each measurement that may or may not be from F is given any weight between 0 and 1. The empirical distribution of just the verified data ($a_k = 0$ for all k) and the empirical distribution of all the data ($a_k = 1$ for all k) are included in $\mathcal{F}(\hat{p})$.

The mle of the set $T(\mathcal{F}(p))$ for unrestricted v is the set of $T(\hat{F})$ with $\hat{F} \in \mathcal{F}(\hat{p})$. Each mle $T(\hat{F})$ can be written as $\sum \phi(t_k)\hat{f}_k$ and lies in the interval $[\hat{T}_L, \hat{T}_U]$, where $\hat{T}_L$ and $\hat{T}_U$ solve

$$\begin{aligned} \sum \psi_L(\phi(X_i) - \hat{T}_L, Y_i) &= 0 \\ \sum \psi_U(\phi(X_i) - \hat{T}_U, Y_i) &= 0, \end{aligned}$$

for ψ_L and ψ_U as given in Theorem 1. The endpoints $\hat{T}_L, \hat{T}_U$ are weighted averages of the data. For example, $\hat{T}_L$ gives each verified measurement and each unverified measurement smaller than $\hat{T}_L$ weight one and each larger unverified measurement weight zero.

The endpoints $\hat{T}_L$ and $\hat{T}_U$ of the mle interval are M-estimators. Under appropriate moment assumptions, standard theory for M-estimators (Hampel, Ronchetti, and Rousseeuw, 1986) implies that $\hat{T}_L$ is asymptotically normal(T_L, σ_L^2/n) where

$$\sigma_L^2 = \frac{\int \psi_L(\phi(x) - T_L, 1)^2 p(dx, 1) + \int \psi_L(\phi(x) - T_L, 0)^2 p(dx, 0)}{\left(\int \psi_L'(\phi(x) - T_L, 1) p(dx, 1) + \int \psi_L'(\phi(x) - T_L, 0) p(dx, 0)\right)^2}$$

and $\psi_L'(a, y)$ is the first derivative of ψ_L with respect to a. The usual estimator of σ_L^2 is

$$s_L^2 = \frac{n^{-1} \sum_i \psi_L(\phi(X_i) - \hat{T}_L, Y_i)^2}{(1 - N_L/n)^2},$$

where N_L is the number of unverified measurements with ϕ values larger than $\hat{T}_L$. The asymptotic distribution and standard error of $\hat{T}_U$ are analogous; $\psi_L(a, y)$ is replaced by $\psi_U(a, y)$ and N_L is replaced by the number N_U of unverified measurements with ϕ values smaller than $\hat{T}_U$. The interval

$$[\hat{T}_L - \sqrt{n} z_{\alpha/2} s_L,\ \hat{T}_U + \sqrt{n} z_{\alpha/2} s_U] \tag{4}$$

is an approximate $100(1-\alpha)\%$ confidence interval for $T(F)$, in the sense that the probability that it covers the interval $[T_L, T_U]$ is approximately $1 - \alpha$. If n is not large and F is far from normal, then bootstrapping $\hat{T}_L$ and $\hat{T}_U$ may give better intervals (Efron and Tibshirani, 1993).

Now suppose v is restricted to be nondecreasing. The mle of $\mathcal{F}^*(p)$ is the set $\mathcal{F}^*(\hat{p})$ of distributions on $\{t_1, \ldots, t_K\}$ that correspond to nondecreasing values of $\hat{v}_k$. Because $\hat{v}_k = 1$ for all

k is nondecreasing and satisfies inequality (3), the empirical distribution $\hat{F}^*$ of the verified measurements, which assigns mass $\hat{f}_k^* = n_{k1}/n_1$ to t_k, belongs to the set $\mathcal{F}^*(\hat{p})$. But if no unverified measurement is tied with the smallest verified measurement, say t_{k_0}, then this $\hat{F}^*$ is the only member of the restricted mle set $\mathcal{F}^*(\hat{p})$. To see this, note that any mle of v for the restricted problem satisfies inequality (3). Thus, any mle of v is 1 at t_{k_0} and hence, by monotonicity, 1 at all larger t_k. Therefore, $\hat{F}^*$ is the only member of $\mathcal{F}^*(\hat{p})$ and the mle $T(\mathcal{F}^*(\hat{p}))$ of the restricted set of possible functional values $T(\mathcal{F}^*(p))$ consists of just the single value

$$\hat{T}_U^* = \frac{\sum \phi(X_i) Y_i}{\sum Y_i},$$

which is the average over the verified measurements. This estimator is consistent for the upper endpoint T_U^* of $T(\mathcal{F}^*(p)) = [T_L^*, T_U^*]$. It is also asymptotically normal with mean T_U^* and variance σ_1^2/n_1 where σ_1^2 is estimated by $\sum (X_i - \hat{T}_U^*)^2 Y_i / n_1$.

To summarize, standard nonparametric maximum likelihood does not provide an estimate of the lower limit T_L^* on possible values of $T(F)$ when v and ϕ are nondecreasing and no unverified measurement is tied with the smallest verified measurement. An *ad hoc* estimator $\tilde{T}_L^*$ can be obtained by substituting a density estimate $\tilde{p}$ for p in the definition of q, substituting the estimated $\tilde{q}$ into ψ_L to obtain a $\tilde{\psi}_L$, and defining $\tilde{T}_L^*$ as the solution to

$$\sum \tilde{\psi}_L^*(X_i, \tilde{T}_L^*) Y_i = 0.$$

The estimate $\tilde{T}_L^*$ is a weighted mean of the verified data, where the weights depend on the unverified data through $\tilde{q}$. The large sample behavior of this estimator and how it depends on the choice of density estimate is not obvious; bootstrapping might provide useful approximate standard errors and confidence intervals.

5 An Example

The 1988 Love Canal study compared the concentrations of certain pollutants in a ring around Love Canal to those in another region of upstate New York that had similar soil conditions, population density, and demographics but was not close to a known waste dump. Previous studies had concluded that there was no spatial pattern to the concentrations in the Love Canal ring, so only the levels of concentrations in the two regions were of interest. No additional structure was imposed or of interest. Here we compare mean concentrations in the two regions for the pollutants α-BHC and 2-CNAP. Our methods apply to geometric means (or means of log concentrations) as well.

Six laboratories participated in the study. Soil samples were taken at random locations from the Love Canal and Comparison regions and assigned to the six laboratories by a randomization scheme that balanced the number of samples from each region across laboratories. Laboratories were not told which samples came from which region and had to agree to rigid, unusually exacting analytical protocols and quality control procedures to participate in this study; details are given in CH2M-Hill (1988). Measurements within laboratories and regions were considered to be independent and identically distributed for each compound. Samples across laboratories were also considered to be independent. Table 1 summarizes the data for α-BHC and 2-CNAP by laboratory.

Table 1: Characteristics of the 1988 Love Canal Data on α-BHC and 2-Chloronapthalene (2-CNAP) by laboratory. n is the number of measurements, p_0 the fraction of unverified measurements, p_{cen} the fraction of measurements reported as "below .2 ppb", and $\bar{X}_1$ the mean of the verified measurements. All censored measurements were unverified. The laboratory numbers are those used in CH2M-Hill (1988).

	Canal							Comparison						
		α-BHC			2-CNAP				α-BHC			2-CNAP		
Lab	n	p_0	p_{cen}	$\bar{X}_1$	p_0	p_{cen}	$\bar{X}_1$	n	p_0	p_{cen}	$\bar{X}_1$	p_0	p_{cen}	$\bar{X}_1$
1	89	.13	.00	2.74	.06	.00	.123	29	.34	.00	.55	.03	.00	.096
2	93	.31	.00	1.66	.23	.00	.065	27	.56	.00	.10	.33	.00	.072
3	93	.29	.04	3.65	.68	.03	.084	31	.52	.00	1.93	.58	.00	.065
6	96	.18	.04	4.43	.32	.02	.049	36	.39	.33	.39	.33	.08	.044
7	86	.22	.01	2.15	.09	.03	.086	27	.52	.22	2.97	.07	.04	.100
8	104	.30	.00	6.29	.29	.00	.112	34	.71	.00	.18	.18	.00	.099

As often happens, the fractions of unverified measurements (called nondetects in this and most environmental studies) vary greatly across laboratories and regions. Some unverified measurements are censored, because the analytical protocols allowed unverified measurements below .2 ppb to be reported as "less than .2 ppb." Not all measurements below .2 ppb are censored, however. For example, the smallest α-BHC measurement reported numerically (but not necessarily verified) ranges from .0005 ppb for laboratory 3 to .022 ppb for laboratory 6. The laboratories in this study were not considered unusually diverse. Here, we take the censored measurements to be zero, but replacing them with the smallest numerical measurement of a pollutant in a laboratory changes our estimates only slightly. Substituting the official censoring limit of .2 ppb instead increases the mean estimates noticeably, however, because so many measurements are below .2 ppb.

5.1 Estimated Bounds on the Probability of Verification

To start, estimate a separate $v(x)$ for each region and laboratory pair. Inequality (1) implies that each v satisfies $p(dx, 1)/p_X(dx) \leq v(x) \leq 1$. The lower bound is estimated by taking a density estimate $\tilde{p}_{X|V}(x)$ of the *logged* verified data, a density estimate $\tilde{p}_{X|U}(x)$ of the logged unverified data, and letting $\tilde{p}_X(x) = (n_1/n)\tilde{p}_{X|V}(x) + (1 - n_1/n)\tilde{p}_{X|U}(x)$. Here we computed a normal kernel density estimator in the statistical language S with a width of 4; censored observations were replaced by the smallest uncensored measurement for the same compound and laboratory to avoid taking logs of zero. Other density estimates are possible; a kernel estimate was chosen for convenience.) The results for laboratories 6, 7 and 8 are shown in Figure 2; the plots for the other laboratories are similar.

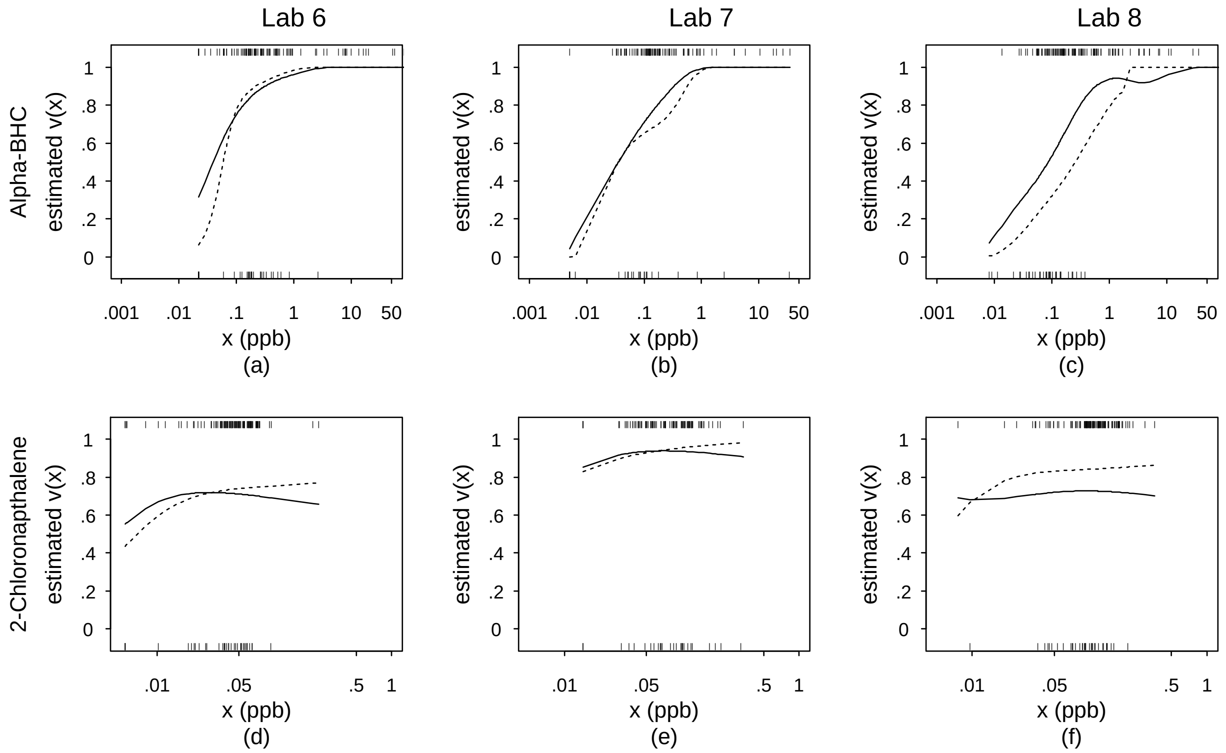


Figure 2: The estimated, unrestricted lower bound $\hat{v}(x)$ for the Canal (solid line) and Comparison regions (dotted line) for the laboratories and compounds shown in Figure 1. The estimated lower bounds are nearly nondecreasing. The points along the bottom and top axes in each plot denote the Comparison and Canal measurements respectively.

For each laboratory, the estimated lower bound for 2-CNAP is nearly a constant over the range of measurements obtained in this study. That constant differs among laboratories and sometimes between regions within a laboratory, however. The latter difference is not surprising. The probability of verification depends not only on the concentration of the pollutant but also on the concentration of other compounds in the soil sample. If confusable compounds are more prevalent in the Canal samples than in the Comparison samples, then a laboratory may have a lower probability of verification in the Canal region than in the Comparison region.

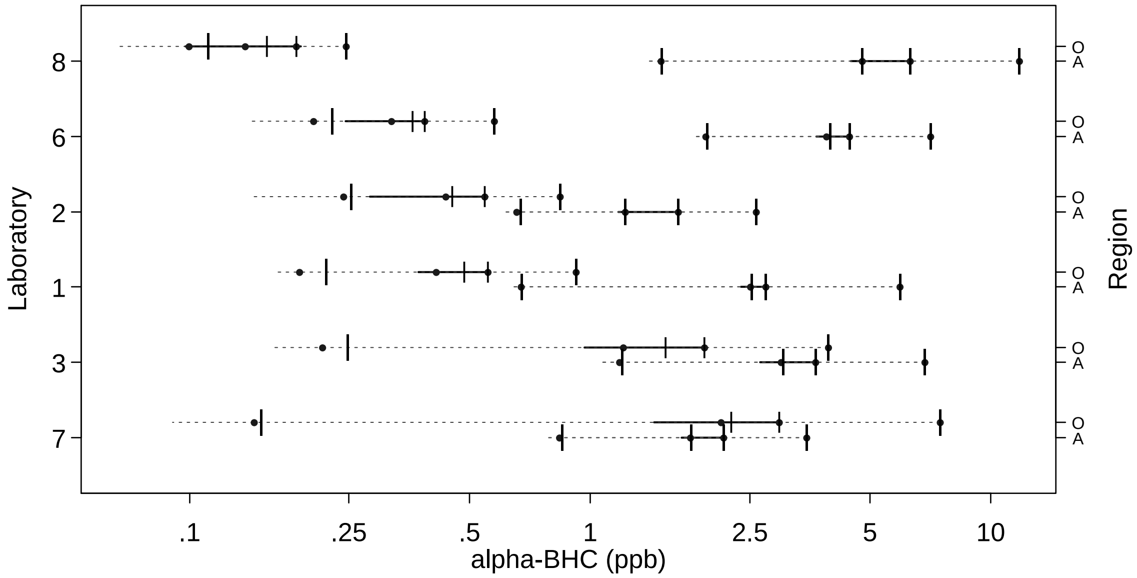
For α-BHC, the estimated lower bounds on $v(x)$ have similar shapes in the Comparison and Canal regions, but the values at any particular x can differ. For laboratory 6, the difference in the estimated lower limits for the two regions is largest for measurements below .1 ppb; for laboratory 7 the difference is largest between .1 and 1.0 ppb, and for laboratory 8 the two curves appear to be location translations of each other on the log concentration scale. Thus, allowing the two regions to have different probabilities of verification is not unreasonable.

5.2 Unrestricted MLEs of Mean Concentration

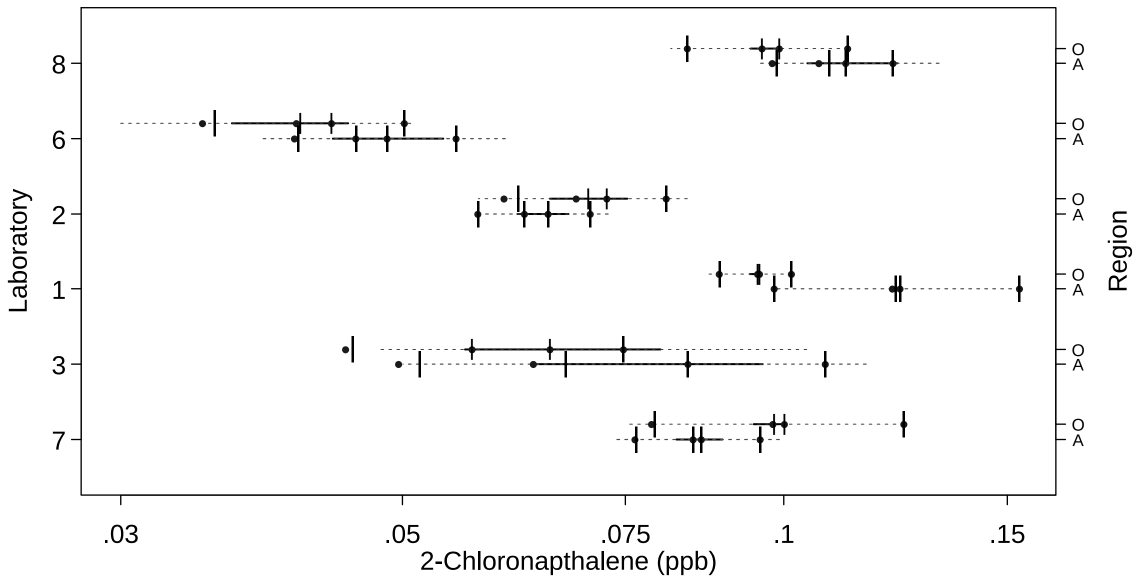
Consider each laboratory separately. If the probability of verification for a laboratory is unrestricted, then $n_{k1}/n_{k+} \leq \hat{v}_k \leq 1$ at its observed t_k's. If the laboratory has no ties across regions, then its lower bounds on $\hat{v}_k$ for the Canal and Comparison regions can be specified separately, even if the regions are assumed to share the same $v(x)$. This implies that the mle of the interval of Comparison means and the mle of the interval of Canal means do not depend on whether the same or different $v(x)$'s holds for the two regions. Similarly, if all t_k that are observed in both regions are either all verified or all unverified, then the the mle intervals do not depend on whether the two regions have the same or different probabilities of verification.

In our data, all α-BHC ties between regions are unverified. For 2-CNAP, laboratory 1 has no ties across regions and laboratories 3 and 6 have only ties in unverified measurements. Laboratories 7 and 8 have one measurement each that is verified in the Canal but not the Comparison region. Moving $\hat{v}$ between 0 and 1 at these two ties has negligible effect on the estimated means of F. Therefore, we allow each laboratory to have a different $v(x)$ in each region.

Figure 3 shows the estimated intervals of pollutant means for each laboratory and compound. For α-BHC, the mle interval for the Canal region is completely to the right of the mle interval for the Comparison region in all but laboratory 7, suggesting that the mean concentration of α-BHC is higher in the Canal region than in the Comparison region. For 2-CNAP, between laboratory variability dominates within laboratory variability. The 2-CNAP mle intervals for the two regions overlap greatly in laboratory 3, slightly in laboratories 2 and 6, and are otherwise disjoint. In laboratory 7, the mle interval for the Canal is to the left of the mle interval for the Comparison region. Bootstrap confidence intervals for each laboratory are also shown in Figure 3. (To bootstrap, random samples with replacement of size n were drawn from the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ 2000 times.) Most of the confidence intervals for the Canal and Comparison regions are wide and overlap. The only exceptions are the confidence intervals for α-BHC in laboratories 6 and 8. The confidence intervals given by equation 4 have negative lower bounds and smaller upper bounds, so are not shown.



(a)



(b)

Figure 3: Estimates of the interval of means of measurements on each of two pollutants and six laboratories in the Love Canal (A) and Comparison (O) regions. A thick line segment shows the unrestricted nonparametric mle $[\hat{T}_L, \hat{T}_U]$. The light extensions of these segments show unrestricted bootstrap 90% confidence intervals for $[T_L, T_U]$. The inner pair of fat dots on a segment show the ends of the restricted estimate $[\hat{T}^*_L, \hat{T}^*_U]$ obtained by assuming each laboratory has a different, nondecreasing v in each region; the outer pair of fat dots show the ends of the corresponding bootstrap 90% confidence interval. The vertical bars show the ends of the restricted estimated intervals assuming that both regions have the same nondecreasing v; the outer vertical bars show the corresponding bootstrap 90% confidence interval. All confidence intervals are based on 2000 bootstrap samples.

5.3 MLEs of Mean Concentration Assuming v is Nondecreasing

First, suppose that each laboratory has different nondecreasing verification functions in the Canal and Comparison regions. This assumption is reasonable for measurements from laboratory 8, in particular (see Figure 2). Because $n_{k1} = 1$ at the smallest verified t_k for each laboratory and region, the right endpoint of each mle interval is the mean of the verified data. The left endpoint is estimated using a density estimate for p, as in Section 5.1. The mle intervals and corresponding bootstrap confidence intervals are shown as fat dots in Figure 3. The restricted estimates and intervals are sometimes but not always noticeably shorter than their unrestricted counterparts, especially for α-BHC in the Canal region. The restricted confidence intervals for α-BHC in the Comparison region are slightly shorter than the unrestricted confidence intervals, however, especially for laboratories 2, 6, and 8.

Finally, suppose that the Canal and Comparison regions have the same $v(x)$ for a given laboratory, but different laboratories have a different $v(x)$. The lower limit on the shared $v(x)$ is estimated by computing $\tilde{p}(x,1)/\tilde{p}_X(x)$ for each region and laboratory separately, using kernel density estimates, and defining $\tilde{q}(x) = max(\tilde{q}_A(x), \tilde{q}_O(x))$, where the subscript O denotes the Comparison region and A the Canal region. The estimated $\tilde{q}(x)$ is then substituted into the definition of $\tilde{T}_L^*$. The resulting sets of estimates and the corresponding confidence intervals are shown as bars in Figure 3. In most cases, the intervals obtained assuming the same nondecreasing $v(x)$ holds for both regions are nearly identical to the intervals obtained assuming that each region has its own nondecreasing $v(x)$.

In summary, treating the data from each laboratory separately, the assumption that $v(x)$ is nondecreasing shortens the confidence intervals for the mean, but not dramatically. Further assuming that the probability of verification is the same for the two regions has little additional effect.

5.4 Estimated Difference in Canal and Comparison Mean Measurements

The obvious question is whether, taken as a whole, the data suggest that the mean measurement for a pollutant is higher in the ring around Love Canal than in the Comparison region. An *ad hoc* answer can be based on the estimated sets of means for each region and laboratory as follows.

Take $\hat{\theta}_{iU} = \hat{T}_{iU,A} - \hat{T}_{iL,O}$ as an estimated upper bound on the interval of the difference in means of the Canal (A) and Comparison (O) regions for laboratory i. Take $\hat{\theta}_{iL} = \hat{T}_{iL,A} - \hat{T}_{iU,O}$ as an estimated lower bound on the interval of difference in means for laboratory i. Roughly speaking, the variability of $\hat{\theta}_{iU}$ should be proportional to $w_i^{-1} = (\tilde{n}_A^{-1} + \tilde{n}_O^{-1})$, where $\tilde{n}_A = \sum_k b_k$ is the sum of the weights b_k on the measurements t_k used to compute $T_{iU,A}$, for example. An *ad hoc* estimate of the interval of difference in means that combines the information from all laboratories has endpoints

$$\hat{\theta}_U = \frac{\sum_i w_i \theta_{iU}}{\sum_i w_i}, \quad \hat{\theta}_L = \frac{\sum_i w_i \theta_{iL}}{\sum_i w_i}.$$

The estimated intervals based on the unrestricted mean estimates, the restricted mean estimates assuming different nondecreasing $v(x)$ in each region, and the the restricted mean estimates assuming the same nondecreasing $v(x)$ in the two regions are given in Table 2. The table suggests that the difference in mean 2-CNAP concentration is not statistically significant, but α-BHC has a higher mean in the Canal region than in the Comparison region. The confidence intervals for α-BHC become tighter as more is assumed about the probabilities of verification in the two regions, but that is not true for 2-CNAP.

Table 2: Estimated difference in means for the Canal and Comparison regions. The interval of estimates describes the differences in the means of the Canal and Comparison distributions that can generate the observed data. The confidence limits are the endpoints of 90% confidence intervals obtained by bootstrapping.

Assumptions	interval of estimates				confidence limits			
	α-BHC		CNAP		α-BHC		CNAP	
unrestricted	1.71	2.99	-.001	.018	.57	4.26	-.009	.026
nondecreasing v in each region	1.84	2.83	.003	.010	.68	4.13	-.006	.019
same nondecreasing v in both regions	1.87	2.64	.004	.010	.71	3.91	-.005	.018
all measurements from F verified	2.43		.007		1.15	3.75	-.001	.016.

The last row of Table 2 shows the estimated differences in mean concentration based on the mean of the verified data. These estimates would be unbiased if all measurements from F were verified ($v(x) = 1$ for all x) or if all measurements from F were equally likely to be verified.

6 Discussion

In our statistical model, some observations are from a nuisance distribution G rather than from the distribution F of interest and there is an imperfect test for verifying that an observation is from F instead of G. Unless strong assumptions are imposed, the distribution F is not identifiable. Instead, given the distribution p of the observables, there is a corresponding set $\mathcal{F}(p)$ of distributions F that can generate p. Not all F belong to $\mathcal{F}(p)$, however, and certain features of F, such as its mean, are restricted to intervals. The length of an interval depends on the probability π that a measurement is from F and on how much is assumed about F, G and v. If π is small and nothing is assumed about (F, G, v), then p has little information about F and the corresponding interval of possible means of F is large. We have not explored the exact relationship between π and the length of the interval of means of F. The interval of means corresponding to $\mathcal{F}(p)$ can be estimated, however, and may be short enough to be useful for inference, as our example shows.

Our model for samples with unverified data has a curious theoretical property. When no restrictions are imposed on (F, G, v), maximizing the nonparametric likelihood produces a set of estimates that converges to $\mathcal{F}(p)$. But when v is restricted to be nondecreasing, maximizing the nonparametric likelihood produces only a single estimate that converges to just one of the distributions F in $\mathcal{F}^*(p)$. Of course, that one limiting F may or may not be the one of interest.

The sample mean $\bar{X}_1$ of the verified data is a common estimate of the mean of the distribution F of interest in environmental applications. If the probability of verifying a measurement from F is one, then $\bar{X}_1$ is the unique mle of $T(F)$ and is unbiased. If the probability of verification is less than one but constant, then $\bar{X}_1$ is still unbiased but possibly inefficient, because the unverified data may also have information about $T(F)$. If v is nondecreasing but otherwise unrestricted, $\bar{X}_1$ is an mle of the largest possible mean over $\mathcal{F}(p)$. Whether estimating just the largest possible mean is satisfactory depends on how long the interval of possible means is. The mle of the estimated interval of possible means, which is easy to compute, suggests how much can be lost by using the mean of the verified data to estimate F. Thus, the mle intervals for arbitrary v and nondecreasing v suggest the robustness of the estimated mean to the assumption that all measurements from F are verified. Of course, the sensitivity analysis is possible only if the statistician has the numerical values of the unverified measurements. Because the instrumentation records these values, they should often be possible to obtain.

Finally, our example shows that the estimated interval of means over $\mathcal{F}(p)$ can be small enough to be useful. Many applications have more structure and heavily censored data than ours, though, and the way that we handled these complications in our example may be inappropriate if there is more structure or censoring. Extensions of our results that allow covariates and censored data would be welcome.

APPENDIX: PROOFS OF THEOREMS

Proof of Theorem 1 Let $H_L(c) = E[\psi_L(\phi(X) - c, Y)]$. Then H_L is continuous and strictly decreasing, because for any $c < c'$

$$H_L(c') - H_L(c) \leq E[(c - c')Y] = (c - c')P(Y = 1) < 0.$$

Because $\lim_{c \to -\infty} H_L(c) = \infty$ and $\lim_{c \to \infty} H_L(c) = -\infty$, there is a unique solution to $H_L(c) = 0$. Let T_L denote that solution. Then v_L satisfies (1). If v is any other function that satisfies (1) and

F has a density with respect to $p(dx, 1)$ that is proportional to $1/v(x)$, then

$$\begin{aligned}
T(F) - T_L &= \frac{\int(\phi(x) - T_L)/v(x)p(dx,1)}{\int 1/v(x)p(dx,1)} \\
&\geq \frac{\int(\phi(x) - T_L)/v_L(x)p(dx,1)}{\int 1/v(x)p(dx,1)} \\
&= \frac{H_L(T_L)}{\int 1/v(x)p(dx,1)} \\
&= 0
\end{aligned}$$

with equality when $v = v_L$. Thus, T_L is a lower bound on the possible values of $T(F)$, and this lower bound is attained. The argument for T_U is analogous. Further, any convex combination $\alpha v_L + (1-\alpha)v_U$ for $0 \leq \alpha \leq 1$ also satisfies (1) and the corresponding $T(F;\alpha)$ varies continuously with α. Therefore, any value in the interval $[T_L, T_U]$ is attained by an F that generates p.

Proof of Theorem 2 If v is nondecreasing, then inequality (1) can be narrowed to

$$q(x) \leq v(x) \leq 1. \tag{5}$$

Let v be a nondecreasing function that satisfies inequality (5) and let F be the distribution with a density with respect to $p(dx, 1)$ that is proportional to $1/v(x)$. Then, because ϕ and v are both nondecreasing,

$$\begin{aligned}
T_U^* - T(F) &= \frac{\int \phi(x)v(x)F(dx)}{\int v(x)F(dx)} - T(F) \\
&= \frac{\int(\phi(x) - T(F))v(x)F(dx)}{\int v(x)F(dx)} \\
&\geq \frac{v(T(F))\int(\phi(x) - T(F))F(dx)}{\int v(x)F(dx)} \\
&= 0.
\end{aligned}$$

Equality is attained by $v_U^*(x) = 1$, which implies that the upper bound is T_U^*. The proof that T_L^* is the unique solution to (2), that T_L^* is a lower bound, and that all values in the interval $[T_L^*, T_U^*]$ are attained is analogous to the proof of Theorem 1.

References

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988), *The New S Language: A Programming Environment for Data Analysis and Graphics*, Pacific Grove, CA: Wadsworth & Brooks.

CH2M-Hill (1988), *Love Canal Emergency Declaration Area Habitability Study, Final Report*, Reston, VA: CH2M-Hill.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm," (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York, N. Y.: Chapman & Hall.

Hampel, F. R., Ronchetti, E. M., and Rousseeuw, P. J. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: Wiley.

Lambert, D., Peterson, B., and Terpenning, I. (1991), "Nondetects, Detection Limits, and the Probability of Detection," *Journal of the American Statistical Association*, 86, 266-277.