Published in final edited form as:

J Agric Biol Environ Stat. 2015 March ; 20(1): 83–99. doi:10.1007/s13253-014-0185-y.

Maximum Pairwise Pseudo-Likelihood Estimation of the Covariance Matrix from Left-Censored Data

Michael P. Jones¹, Sarah S. Perry², and Peter S. Thorne²

¹Department of Biostatistics, University of Iowa, Iowa City, IA 52242

²Department of Occupational and Environmental Health University of Iowa, Iowa City, IA 52242

Abstract

Toxicological studies often depend on laboratory assays that have thresholds below which environmental pollutants cannot be measured with accuracy. Exposure levels below this limit of detection may well be toxic and hence it is vital to use data analytic methods that handle such left-censored data with as little estimation bias as possible. In an on-going study for which our methodology is developed, levels of residential exposure to polychlorinated biphenyls (PCBs) and the interrelationships of their subtypes (congeners) are characterized. In any given sample many of the congeners may fall below the detection limit. The main problem tackled in this paper is estimation of mean exposure levels and corresponding covariance and correlation matrices for a large number of potentially left-censored measures that have very low bias and are computationally feasible. The proposed methods are likelihood based, using marginal likelihoods for means and variances and pairwise pseudo-likelihoods for correlations and covariances. In the simple bi-variate case, head-to-head comparisons show the proposed methods to be computationally more stable than ordinary maximum likelihood estimates (MLEs) and still maintain comparable bias. When the number of variables is much larger than 2, the proposed methods are far more computationally feasible than MLE. Furthermore, they exhibit much less bias when compared to popular imputation procedures. Analysis of the PCB data uncovered interesting correlational structures.

Keywords

Lower detection limit; Non-detects; Principal component analysis; Polychlorinated biphenyls

1. Introduction

Environmental exposure studies often involve measuring pollutants that can be present at levels below analytical detection, resulting in left-censored data. This frequently occurs when a laboratory's limit of detection is high relative to a typical exposure level, or when the sample volume is small, e.g. due to a short sampling time (Hewett, 2006, p. 415). Furthermore, there may be several limits of detection for a single compound, due to different sample volumes or the use of several laboratories with different levels of detection (Helsel, 2005). Left-censored measurements, also called non-detects, are common when contaminant residues are measured in air, water, soil, food, blood and tissue samples. A few examples include levels of herbicides/pesticides/nitrates in well water (Holden et al., 1992); radiation

measured by film badges worn by radiation workers (Kerr, 1994); toxic metals (e.g. lead, chromium) in highway runoff (Shumway et al., 2002); okadaic acid, arsenic and cadmium in food (EFSA, 2010); and endotoxin levels of environments with organic dust (Thorne et al., 2010).

There are three primary approaches used in computing univariate descriptive statistics, such as the mean and standard deviation. The first are the substitution methods, in which any measure below the limit of detection L is replaced by some value: such as $L/2$ (Nehls and Akland, 1973) or $L/\sqrt{2}$ when L is below the distributional mode (Hornung and Reed, 1990). The usual sample mean and standard deviation formulae are then applied. These methods were originally proposed and are still commonly used for the sake of computational convenience. The second approach is maximum likelihood estimation (MLE) based on a specific parametric distribution, such as the normal (Hald, 1949) or the lognormal; MLEs for left-censored observations on a single variable are readily available in statistical software packages, e.g. the `survreg` function in R (2005) and the `Lifereg` procedure in SAS (SAS Institute, 2004). The third approach is log probit regression, also known as regression on order statistics (Hewett, 2006, p. 417–418). A multitude of papers have investigated the statistical properties of these methods via simulation in terms of bias and mean squared error; see Hewett and Ganser (2007) for a very comprehensive study. Although there is no one method that is best in all settings, as categorized by sample size and amount of censoring, the MLE is generally preferred. A serious complication can arise when the data have been transformed before computing the MLEs, and one is interested in estimating means on the original scale. Transformation bias can occur when back-transforming the MLEs to the original scale. Shumway et al. (1989, 2002) have developed the Box-Cox transform for left-censored data and bias correction for back-transformed estimates. However, in this paper we are interested in parameter estimates on the transformed scale and therefore will not consider back transformation issues.

For the bivariate problem, Lyles et al. (2001) used maximum likelihood based on a bivariate normal distribution to simultaneously estimate the means, variances and correlation between two left-censored log-transformed variables. The goal in this paper is to develop likelihood-based methods for estimating the mean vector and covariance and correlation matrices from left-censored data for the p -variate problem ($p \geq 2$). As discussed in the next section, there are up to 2^p possible censoring patterns, and hence simultaneous parameter estimation via maximizing the resulting complicated likelihood is not feasible. In Section 2 we propose a method of using maximum marginal likelihood estimates of the means and variances and maximum pairwise pseudo-likelihood estimates of the correlations (and covariances). Section 3 contains a simulation study comparing the proposed methods with ones commonly used in practice. The research methods proposed in this paper have been developed in conjunction with a large on-going NIH-funded study measuring the levels of 159 polychlorinated biphenyl (PCB) congener groupings from human blood samples. In addition to the basic descriptive statistics of means and variances, the correlational structure among the congener groupings is of real interest, as is the potential simplification to a few principal components. The large number of measures and the high level of left censoring represent a serious challenge that has motivated our work. A full description of the problem and an

analysis of the year-two study data is given in Section 4. Our overall conclusions are reported in Section 5.

2. Methods

In this section we describe methods for estimating the vector of means and covariance matrix from normally distributed data subject to left censoring. Specifically, we use maximum marginal likelihood estimates for the means and variances and propose maximum pairwise pseudo-likelihood estimation of the correlations. We finish this section by discussing several alternative approaches.

Let $\mathbf{Y}_i^\circ = (Y_{i1}^\circ, \dots, Y_{ip}^\circ)'$ be the uncensored measurements for the, i^{th} unit or subject. In our motivating example these are the true (uncensored) levels of p different PCB congeners in the i^{th} sample. Furthermore, let $\mathbf{L}_i^\circ = (L_{i1}^\circ, \dots, L_{ip}^\circ)'$ be the lower limits of detection for the i^{th} sample. This level of generality is desirable since there may be several detection limits for the same compound; e.g. due to combining data from different labs. We assume that the data have been properly transformed, if necessary, so that for all (j, k)

$$(Y_{1j}^\circ, Y_{1k}^\circ), \dots, (Y_{nj}^\circ, Y_{nk}^\circ) \sim iid N_2(\mu_j, \mu_k, \sigma_j^2, \sigma_k^2, \rho_{jk}) \quad (2.1)$$

and that L_{ij}° and Y_{ij}° are independent. Note that full multivariate normality is not assumed for $\mathbf{Y}^\circ$. Next, let $Y_{ij} = \max\{Y_{ij}^\circ, L_{ij}^\circ\}$ and $\eta_{ij} = 1[Y_{ij}^\circ \geq L_{ij}^\circ]$, so that Y_{ij} is left-censored when η_{ij} is 0 and is uncensored when η_{ij} is 1.

Our goal is to estimate the vector of means $\boldsymbol{\mu}$ and covariance matrix from the observed data $\{(Y_{ij}, \eta_{ij}) : j = 1, \dots, p; i = 1, \dots, n\}$. We propose a simple approach, which is especially appealing when p is large. This approach has two steps, listed next with details to follow.

- Step 1 : Obtain the maximum marginal likelihood estimates (MMLEs) of the j^{th} mean and variance $(\hat{\mu}_j, \hat{\sigma}_j^2)$ and their standard errors from the data subset $\{(Y_{ij}, \eta_{ij}) : i = 1, \dots, n\}$. This is done separately for each $j, j = 1, \dots, p$ hence the term marginal.
- Step 2: Let $L_{jk}(\mu_j, \mu_k, \sigma_j^2, \sigma_k^2, \rho_{jk})$ be the left-censored bivariate normal likelihood based on the observations $\{(Y_{ij}, \eta_{ij}, Y_{ik}, \eta_{ik}) : i = 1, \dots, n\}$. Define the pair-specific pseudo-likelihood

$$PPL_{jk}(\rho_{jk}) = L_{jk}(\hat{\mu}_j, \hat{\mu}_k, \hat{\sigma}_j^2, \hat{\sigma}_k^2, \rho_{jk}) \quad (2.2)$$

using $\hat{\mu}$'s and $\hat{\sigma}^2$'s from the marginal estimation in Step 1. The pseudo-likelihood (2.2) is maximized to find $\hat{\rho}_{jk}$ the estimated correlation between Y_{ij}° and Y_{ik}° . This maximization is repeated for all pairs j, k . The estimated covariance matrix is derived from the estimated variances and correlation matrix. Bootstrap resampling can be used to estimate confidence intervals of the Fisher-Z transformed correlations, which are then back-transformed to obtain confidence intervals for the correlations.

Details for Step 1. The j^{th} marginal likelihood

$$L_j(\mu_j, \sigma_j^2) = \prod_{i=1}^n \left[f(y_{ij}; \mu_j, \sigma_j^2) \right]^{\eta_{ij}} \left[F(y_{ij}; \mu_j, \sigma_j^2) \right]^{1-\eta_{ij}}$$

is maximized to find $(\hat{\mu}_j, \hat{\sigma}_j^2)$, where f and F are the normal density and distribution function, respectively. These MMLs are consistent but not necessarily efficient since other correlated Y 's may contain information. Standard errors (se's) are derived from the negative inverse Hessian matrix as part of the Newton-Raphson algorithm. The Wald 95% confidence interval for μ_j is $\hat{\mu}_j \pm 1.96se(\hat{\mu}_j)$. The MMLs and standard errors are computed by the `survreg` function in R (2005) or `Proc Lifereg` in SAS (SAS Institute, 2004). For example, in R one would use the statement

`survreg(Surv (Y[j], eta[, j])~1, dist="gaussian", type="left").`

Details for Step 2. There are several functions needed in the likelihood construction. The bivariate normal density of $(Y_{ij}^\circ, Y_{ik}^\circ)$ is

$$f_{jk}(a, b) = \frac{1}{2\pi\sigma_j\sigma_k\sqrt{1-\rho_{jk}^2}} \times \exp \left\{ \frac{-1}{2(1-\rho_{jk}^2)} \left[\left(\frac{a-\mu_j}{\sigma_j} \right)^2 - 2\rho_{jk} \left(\frac{a-\mu_j}{\sigma_j} \right) \left(\frac{b-\mu_k}{\sigma_k} \right) + \left(\frac{b-\mu_k}{\sigma_k} \right)^2 \right] \right\}.$$

The bivariate normal cumulative distribution function is

$$F_{jk}(a, b) = \int_{-\infty}^a \int_{-\infty}^b f_{jk}(s, t) ds dt,$$

and is readily computed in R using the `mvtnorm` library (Genz et al., 2011) which is based on an algorithm by Genz (1992). It is easily shown that

$$(\partial/\partial a)F_{jk}(a, b) = \int_{-\infty}^b f_{jk}(a, t) dt = \frac{1}{\sqrt{2\pi}\sigma_j} \exp \left\{ -\frac{1}{2} \left(\frac{a-\mu_j}{\sigma_j} \right)^2 \right\} \Phi \left(\left[\frac{b-\mu_k}{\sigma_k} - \rho_{jk} \frac{a-\mu_j}{\sigma_j} \right] / \sqrt{1-\rho_{jk}^2} \right)$$

where Φ is the standard univariate $N(0,1)$ distribution function. The derivative $(\partial/\partial b)F_{jk}(a, b)$ is defined similarly, interchanging the roles of a and b , j and k .

The bivariate normal likelihood $L_{jk}(\mu_j, \mu_k, \sigma_j^2, \sigma_k^2, \rho_{jk})$ for left-censored data is

$$\prod_{i=1}^n [f_{jk}(y_{ij}, y_{ik})]^{\eta_{ij}\eta_{ik}} \left[(\partial/\partial a)F_{jk}(a, y_{ik}) \right]_{a=y_{ij}}^{\eta_{ij}(1-\eta_{ik})} \times \left[(\partial/\partial b)F_{jk}(y_{ij}, b) \right]_{b=y_{ik}}^{(1-\eta_{ij})\eta_{ik}} [F_{jk}(y_{ij}, y_{ik})]^{(1-\eta_{ij})(1-\eta_{ik})}. \quad (2.3)$$

Since the MMLEs, $\hat{\mu}_j$, $\hat{\mu}_k$, $\hat{\sigma}_j^2$ and $\hat{\sigma}_k^2$ from Step 1, are $\sqrt{n}$ — consistent estimators, we employ a pseudo-likelihood approach, substituting these estimators into (2.3). The resulting pair-specific pseudo-likelihood $PPL_{jk}(\rho_{jk}) = L_{jk}(\hat{\mu}_j, \hat{\mu}_k, \hat{\sigma}_j^2, \hat{\sigma}_k^2, \rho_{jk})$ is maximized to provide a maximum pairwise pseudo-likelihood estimator (MPPLE) $\hat{\rho}_{jk}$, which is consistent by pseudo-likelihood theory (Gong and Samaniego, 1981). The MPPLE of the covariance is $\hat{\sigma}_{jk} = \hat{\sigma}_j \hat{\sigma}_k \hat{\rho}_{jk}$. This is repeated for all $j < k$. Conceptually, our proposal, using (2.2), is a pairwise pseudo-likelihood

$$PPL = \prod_{j < k} PPL_{jk}(\rho_{jk}), \quad (2.4)$$

which allows each ρ_{jk} to be estimated separately from the pair-specific pseudo-likelihoods.

Maximization of (2.2) is easily implemented. Since there is only one parameter ρ_{jk} , which lies in $[-1, 1]$, one can simply employ a linear search, such as a Newton algorithm based on numerical derivatives of the gradient and the Hessian. A popular implementation is the `nlm` function in R. The Newton-Raphson algorithm is less desirable because it requires analytic derivatives of (2.2).

Since MPPLE computation is fast, bootstrap resamples can be taken of the original data to provide standard error estimates, $se_B(\hat{g}_{jk})$, of the Fisher-Z transformed MPPLE

$\hat{g}_{jk} = 0.5 \log\{(1 + \hat{\rho}_{jk})/(1 - \hat{\rho}_{jk})\}$ and the associated Wald-type confidence interval

$\hat{g}_{jk} \pm Z_{1-\alpha/2} \cdot se_B(\hat{g}_{jk})$, where $Z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a standard normal. This interval is then back-transformed via $\rho = \exp(2g - 1)/\exp(2g + 1)$ to attain a bootstrap confidence interval for a correlation. Since the data are bivariate normal, a Fisher-Z transformed estimate of the correlation is approximately normal, justifying the use of the Wald-type bootstrap confidence interval, which requires but a fraction of the number of resamples as needed by the percentile and bias-corrected and accelerated (BC_a) approaches (Efron and Tibshirani, 1993).

There are several likelihood-based alternatives to the proposed two-step approach we considered. The first is classical maximum likelihood estimation (MLE). When $p = 2$, the

full likelihood is given by (2.3) and can be maximized simultaneously over $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Lyles et al. (2001) took this approach, but their interest lay strictly in the bivariate problem. The obvious question is: How well do the proposed MMLEs/MPPLEs compare to the classical MLEs? This is explored in section 3.1 for $p = 2$. For $p > 2$, MLEs rely on the uncensored $\mathbf{Y}^\circ$ data following a multivariate normal distribution, a stronger assumption than the pairwise bivariate normality (2.1) needed for the proposed methodology. More importantly, as the number of variables p increases, the MLE approach becomes less and less feasible. Imagine the likelihood contributions for a multivariate normal when $p = 25$. There is a potential maximum of 2^{25} different combinations of the values of $\{\eta_{i1}, \dots, \eta_{i,25}\}$ for which one might need to compute the likelihood contribution, involving some very complicated integrals of a multivariate normal density. Of course, the actual number is limited by the sample size. Simultaneous maximization over all parameters is not practical. A better option is the EM algorithm which has a simple maximization step but potentially a very large number of difficult conditional expectations for the left-censored observations, again becoming computationally challenging and less attractive. A second alternative is to consider a product of all pairwise bivariate normal likelihoods

$$\prod_{j < k} L_{jk}(\mu_j, \mu_k, \sigma_j^2, \sigma_k^2, \rho_{jk}) \quad (2.5)$$

where L_{jk} is defined by (2.3). Unlike classical maximum likelihood it does not assume full multivariate normality. Superficially, it resembles (2.4) but involves a single stage of estimation. Simultaneous maximization over all parameters with complex censored data makes this approach very computer intensive and hence less appealing. A third alternative is to break (2.5) into $p(p-1)/2$ separate bivariate normal likelihoods and to maximize each separately, rendering a computationally more appealing approach than (2.5). Although it would produce a single estimate for each correlation, there would be $p-1$ estimates of each μ_j and σ_j^2 . The question of which estimator of μ_j and σ_j^2 to choose is not well defined by the “likelihood” itself and is potentially rather ad-hoc. A fourth alternative could start with a multivariate normal likelihood L , as does the classical MLE, but is modified into the pseudo-likelihood $PL(P) = L(\hat{\boldsymbol{\mu}}, \hat{\mathbf{S}}, \mathbf{P})$, using all the data, plugging in the MMLEs $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_p)'$ and $\hat{\mathbf{S}}$, the diagonal matrix with diagonal elements $(\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2)$, and then maximizing to estimate the correlation matrix $\mathbf{P}$. Simultaneous maximization of PL over all $p(p-1)/2$ unique elements of $\mathbf{P}$ is still very computer intensive, especially for large p in the presence of a number of different censoring patterns.

3. Results

In this section we report on several simulation studies, designed to investigate the performance of the proposed methods in terms of bias and square root of the mean squared error (RMSE) for different sample sizes, variances and amounts of left censoring. The single bivariate normal case ($p = 2$) is studied in section 3.1 and the multivariate normal case for p

= 6 in section 3.2. The proposed methods are compared either against classical MLEs ($p = 2$) or against popularly used imputation methods ($p > 2$).

3.1. Bivariate Normal Population

In this subsection we compare the maximum marginal likelihood estimates (MMLEs) and maximum pairwise pseudo-likelihood estimates (MPPLEs) against the classical maximum likelihood estimates (MLEs) when $p = 2$. The MLEs were computed by simultaneously maximizing (2.3) over $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ using the `nlm` function in R. Starting parameter values for the MLE iterative routine were the usual mean, variance and correlation estimators in which the left-censored observations were replaced by imputed values. The `nlm` function returns a code describing the reason for termination of the iteration process. As stated under R's `nlm` description, the codes are: (1) the relative gradient is close to zero, current iterate is probable solution; (2) successive iterates within tolerance, current iterate is probable solution; (3)-(5) codes for which the last iterate is not the solution for different stated reasons. There is, in essence, another code which we call `nlm.code = 0`, in which the `nlm` function results in an error message and cannot proceed further.

In the first set of simulations, the uncensored data were generated from a bivariate normal with means $(\mu_1, \mu_2) = (0, 0)$, variances $(\sigma_1^2, \sigma_2^2) = (2, 3)$ and a fixed correlation ρ . Without loss of generality we chose the true means as $(0, 0)$, so that the average $\hat{\mu}_j$ is also the bias. In all, 16 substudies were conducted, corresponding to combinations of sample sizes of $n = 30$ and $n = 100$, various amounts of left censoring (25%, 50% and 75%) and two correlations, -0.4 and 0.8 . Lower limits of detection (L°) of $-0.9539, 0, 0.9539$ for Y_1 and $-1.1683, 0, 1.1683$ for Y_2 correspond to these three levels of left censoring. The goal was to generate 1000 data sets for which a nonzero `nlm` code resulted. This occurred 100% of the time for the MMLE/MPPLE. We did the same for the MLE, but needed to generate more data sets to achieve the goal of 1000 with nonzero `nlm` codes. Table 1 reports the average values of the MMLEs/MPPLEs and the MLEs based on use of $L/2$ to obtain starting values for those data sets with a `nlm` code of 1 or 2, since they terminate in an apparent solution. It is very important to note that these results are not based on the same data sets and hence this is not a head-to-head comparison but rather a results summary for the methods when convergence is attained. (The head-to-head comparison is described later.) Fewer than 1.5% of the generated data sets result in `nlm` codes of 3 – 5. Note that the biases are typically small and similar between the two methods. The biases are very close for $n = 100$. The most notable bias lies in substudies 3 and 7 (lines 3 and 7 of Table 1), the small-sample cases involving 75% left censoring for both Y_1 and Y_2 , in which both methods produce biased estimates. Substudy 3 had the most serious MLE convergence problems, requiring 1105 data sets to acquire 1000 with a nonzero `nlm` code. Table 2 reports the RMSE values for the same data sets used in Table 1. Most RMSEs are similar between the two methods, except for substudies with $n = 30$ and 75% censoring for Y_1 and Y_2 , where the MLEs are slightly better when $\rho = -0.4$ but the MMLEs of the means and variances are substantially better when $\rho = 0.8$. As a general rule, the lower the censoring rate, the closer the bias and RMSE between the two methods.

Both methods suffer larger bias and RMSE when both Y values undergo 75% censoring. To investigate this further, we took the first 1000 data sets used for substudy 3 in Table 1 and computed the MLEs and MMLEs/MPPLEs. The MLEs were computed with two different starting value approaches: the usual sample estimates based on replacing left-censored values with either $L/2$ or L . All results are summarized by nlm code = 1 and nlm code = 1 or 2 in Table 3. First note that the MMLEs/MPPLEs converged with a nlm code of 1 for all 1000 data sets, exhibiting some level of bias. The MLE results are more complex. Using only those data sets in which the MLE optimization routine resulted in a nlm code of 1 or 2, MLEs with the $L/2$ starting value approach were found for 90.5% of the 1000 data sets. Almost all of the remaining 9.5% of the data sets had a nlm code = 0. Using the stricter criterion of nlm code of 1, MLEs could be found for 86.9% of the data sets. Computing MLEs with the L starting-value approach results in nlm code = 1–2 convergence for 94.3% of the data sets but with enormous bias, whereas using a nlm code = 1, convergence was obtained for only 68.6% of the data sets with much lower bias.

Selection bias enters any comparison when results for one method are only summarized over the “nice” data sets. In order to provide a head-to-head comparison, the MMLEs/MPPLEs were computed over the subset of data sets for which the MLEs converged. These are shown in Table 3 for comparison against the MLE with the $L/2$ starting value approach only, since the L starting value approach is obviously inferior. Note that the biases and RMSEs for the MMLEs/MPPLEs are very similar to the MLEs. Given the poor convergence rates for the MLE routine and how well the MMLEs/MPPLE perform with respect to convergence, bias and RMSE, we tried using these estimators as starting values for the MLE routine. As shown in Table 3, this leads to 99.6% MLE convergence with a nlm code of 1 and 100.0% convergence with a nlm code of 1–2, and very importantly, with almost identical biases and RMSEs as the MMLE/MPPLEs. In summary of this highly-censored setting, the MLE convergence problem stems from choice of starting values. The best approach is to use the MMLEs/MPPLE to start the MLE routine. However, the resulting biases and RMSEs for this modified MLE algorithm and the MMLEs/MPPLE are almost identical, implying that both methods work equally well for the bivariate normal setting studied here.

The MMLEs/MPPLEs are computationally much faster and have better convergence rates than the MLEs with reasonably similar bias and RMSE, except for much lower RMSEs in line 7 of Table 2. The question arises as to when the MLE is clearly the preferred method for $p = 2$. Although consistent, the MMLEs are not efficient estimates of μ_2 and σ_2^2 since information from Y_1 is ignored. If Y_1 is lightly censored and Y_2 is heavily censored, then intuitively the MLE should be better able to estimate the mean and variance of Y_2 when the correlation between Y_1 and Y_2 is large. For this reason we studied the special cases of 25% and 75% left censoring of Y_1 and Y_2 , respectively, when $\rho = -0.4$ and 0.8 and report the results in Tables 1 and 2. Here $(\sigma_1^2, \sigma_2^2) = (2, 3)$. Rounding off to the second digit for the sake of comparison, the biases of the MMLEs and MLEs for μ_2 and σ_2^2 (Table 1) are almost identical, except for one case favoring the MMLE. The corresponding RMSEs (Table 2) are quite similar except for one case favoring the MLE ($n = 30$, $\rho = -0.4$) and one case favoring the MMLE ($n = 30$, $\rho = 0.8$). The MMLEs and MLEs of μ_1, σ_1^2 and ρ have almost identical

biases and RMSEs for all (25%, 75%) left-censoring substudies. That the MMLE did so well in this setting is an interesting finding.

We next investigate the effect of varying the variances. Again, if Y_1 is lightly censored (25%) and has small variance (1), then the scales should be tipped in favor of the MLEs of μ_2 and σ_2^2 when Y_2 is heavily censored (75%) with large variance (4). The MLEs tend to have smaller RMSEs for μ_2 and σ_2^2 when $n = 30$, but much of the advantage disappears when $n = 100$. In the case of (75%, 75%) censoring, the MLEs have lower RMSE for μ_2 and σ_2^2 when $\rho = -0.4$, but higher when $\rho = 0.8$, similar to the trend seen in Table 2. Overall, the biases are quite similar between the two methods. The two methods are similar for estimating $(\mu_1, \sigma_1^2, \rho)$. When the censoring rates are both 25% or both 50%, the biases and RMSEs are similar with no consistent advantage to either method (results not shown). Very importantly the biases and RMSEs for estimating the correlations are always similar. Changing the (Y_1, Y_2) variances to (4,1) tends to favor the MLE of σ_1^2 when $\rho = -0.4$ and the MMLE when $\rho = 0.8$ under heavy censoring.

A fourth simulation study (results not shown) investigates the very small sample settings of $n = 15$ and $n = 20$ when (Y_1, Y_2) are either both 25% left censored or both 50% left censored and their variances are (1,4). The biases of the two methods are fairly similar throughout. The RMSEs are similar for (25%,25%) censoring. With (50%,50%) censoring, the MLE has a slightly lower RMSE for (μ_2, σ_2^2) for $n = 15$. The MLE advantage in RMSE disappears for $n = 20$, as it does for $n = 30$ in Table 2.

3.2. Multivariate Normal Population

We now investigate methods of estimating the parameters of the multivariate normal distribution. Maximum marginal likelihood estimates are found for the vectors of means and variances, while maximum pairwise pseudolikelihood estimates are used for the correlation matrix, as described in section 2. For the sake of comparison, we use two common imputation methods.

The vector $(Y_1^\circ, \dots, Y_6^\circ)$ is generated from a multivariate normal with zero means and unit variances, and a correlation matrix consisting of four 3×3 submatrices. The upper left quadrant has all correlations of 0.7, the lower right quadrant has correlations of 0.4 and the remaining correlations are all zero. In this setting the true correlation and covariance matrices are identical.

In the two most commonly used imputation methods, left-censored values are replaced either by $L_j/2$ or by $L_j/\sqrt{2}$ and then treated as if known in the analysis. We refer to the two imputation methods by $L/2$ and $L_j/\sqrt{2}$. The fixed detection limits were chosen so that Y_1° and Y_4° are 25% left censored, Y_2° and Y_5° are 50% left censored, and Y_3° and Y_6° are 75% left censored. Again we assume the original six environmental measures, typically non-negative, have been properly transformed to normality for statistical analysis.

One thousand data sets, each of size $n = 100$, were generated. Table 4 shows that the MMLEs of the 6 means and variances suffer very little bias, whereas the imputation methods suffer increasing overestimation of the mean and dramatic underestimation of the variance as the percentage of left censoring increases. The $L/2$ imputation method has slightly lower RMSE for estimating the mean than the MMLE for (25%, 50%) left censoring, but much larger for 75% left censoring. For estimating the variances, the MMLEs have lower RMSEs except for 25% left censoring. The $L/\sqrt{2}$ imputation method consistently has the worst RMSE values for the means and variances. Table 5 shows that the MPPLEs of the correlations consistently have low bias and are much less biased than the two imputation methods for the non-zero correlations. The RMSEs for the MPPLEs are lower when estimating the non-zero correlations, but are higher when estimating the zero correlations. This is due to that fact that the single imputation methods assume the left-censored measures for the same individual are independent. Confidence interval coverage based on MMLEs and MPPLEs is reasonable in all cases. The $L/2$ -based confidence interval has nominal coverage (not shown) of the mean for 25% censoring, but is poor otherwise. The $L/\sqrt{2}$ has consistently poor coverage of the means. The imputation-based confidence intervals achieved the nominal level for zero correlations but were extremely poor for non-zero correlations.

4. Analysis of Polychlorinated Biphenyl Data

Polychlorinated biphenyls (PCBs) constitute a family of man-made environmental contaminants that are resistant to environmental (chemical and thermal) degradation and are bioaccumulating. They have been implicated in adverse effects on the immune, nervous, endocrine and reproductive systems and are recognized as human carcinogens (Lauby-Secretan et al. 2013). Although most production was banned in 1977 in the USA, they persist in the environment. PCBs were commonly used in hydraulic fluids, sealants, transformers, capacitors, fluorescent light ballasts and paints. As an example, PCBs can constitute a high percentage of window caulk in old buildings, such as schools, and migrate into the surrounding air, dust and soil. Inhalation and ingestion of food, including breast milk, are the primary sources of human and animal exposure.

The family of PCBs consist of 209 different chlorinated biphenyls, called congeners, owing to the availability of 10 chlorine binding sites on the hexagonal-shaped biphenyl rings. The binding patterns determine congener behavior. In general, the lower molecular weight congeners are volatile and hence are more likely to be inhaled, whereas the higher molecular weight congeners are more likely to be ingested. PCB-containing materials constitute complex mixtures of these congeners.

In the on-going NIH-funded AESOP study (Ampleman et al. 2014, Marek et al. 2013), the levels of 209 congeners represented by 159 chromatographic peaks were measured in blood samples on 192 subjects. Measurements were made either of individual congeners, e.g. PCB 66, or co-eluting PCBs, e.g. PCBs 83 and 99. These were then standardized as the number of nanograms (ng) of congener per milliliter (ml) of blood lipids.

There are two key considerations before analyzing the data. The first concerns the concentration measurements as a ratio. The numerator, as the amount of congener, can be

lower than the laboratory detection limit and hence left censored, whereas the denominator, as the amount of lipids in the blood sample, is measured exactly. (The amount of blood taken was the same for all subjects.) Hence, the left-censoring value for the ratio differs for each subject since subjects have different lipid levels but the same lower limit of detection for a congener. This feature of the data will affect any likelihood-based or imputation procedure, but is often ignored in practice.

The second consideration is the normality assumption made by the likelihood procedure. Of the 159 congeners measured, 16 had at least 25% of their values uncensored, a boundary at which the proposed likelihood methods performed reasonably well in the simulation studies. Normal probability plots showed that the congener concentrations were right skewed; however, a cube-root transform resulted in normality, except for three outliers described next. Among the three outliers, two were substantially different for three congeners and the third person was an enormous outlier for 9 congeners, sometimes with concentrations four times as large as the next largest, even though the lipid level was typical. These three subjects are eliminated from further analysis.

Table 6 lists the 16 PCB congeners having sufficient data for analysis. The percentage of uncensored measures ranged from 25.5% to 80.2%. The MMLEs of the means and standard deviations of the cube-root of the PCB concentrations are displayed. Most previously studied cohorts either consumed fish or marine mammals from contaminated waters or were occupationally exposed. The values measured in this study are low in comparison, but are of special interest because of the residential location of its participants. Recall again that the level of a grouped set of congeners are measured by their chromatographic peak and a shared elution time. The PPMLs of the correlations are displayed in Figure 1. In that plot each line traces the correlations of one congener grouping peak with the other congener groupings. For example, peak C (PCB 22) has correlations of 0.66 and 0.57 with peaks A and B, 1.0 with itself, 0.36 with peaks D and E, 0.14 with peak F, etc. The correlational patterns are striking and informative. Figure 1 was split into two frames to better display this. For convenience, we identify the congener groups with the peak labeling in Table 6. Peaks A - G (first panel) have quite different correlational structures from peaks H - P (second panel). Furthermore, within the first set, peak groupings (A,C), (B,E) and (D,F,G) exhibit distinct trajectories with respect to their correlations with peaks H-P. In the second panel peaks H-P have strikingly similar correlational structures. They have generally low correlations with congener peaks A-G and high correlations amongst themselves (H-P). The second panel provides another view of the similarities of their correlational patterns within peak groupings (A,C), (B,E) and (D,F,G).

To better understand the two levels of correlational patterns observed in Figure 1, we need to consider the differences in congener molecular structure. The separation reflects the homolog groupings in which peaks A to G represent di-, tri- tetra-, and penta-chlorobiphenyl congeners and peaks H to P represent hexa-, hepta-, and octa-chlorobiphenyl congeners. The hexa- and higher chlorinated congeners have much lower vapor pressures, higher octanol-water partition coefficients (K_{OW}) and are generally more resistant to metabolism. Moreover, the observed correlational grouping pattern that placed peak A with C, peak B with E and combined peaks D, F and G may reflect similarities in K_{OW} . Specifically, the

higher the $\log(K_{OW})$, the more lipophilic is the compound. PCBs 11 and 22 (peaks A and C) have $\log(K_{OW})$ less than 5.6. PCBs 28 and 66 (peaks B and E) both have $\log(K_{OW}) = 5.8$. PCBs 70/74, 99 and 118 (peaks D, F and G) have values of $\log(K_{OW})$ between 6.0 and 7.0. The rationale for this hypothesis is that differences in K_{OW} affect the uptake, distribution and metabolism in the body as well as their retention in a chromatographic column.

Another goal of the analysis was to conduct a principal component analysis on the 16 congeners. Since all measurements are on the same scale, the PPMLE of the covariance matrix was used. The first two components are given in Table 6. The first component appears as a weighted average of the individual congener groups with all coefficients positive. Interestingly, the second component exhibits a contrast between congener peaks A-G and H-P, confirming the structure seen in the correlation plots. These two components explain almost 80% of the total variability of the analytic data.

5. Discussion

Measurements falling below the lower limit of detection are a common and an important problem in environmental data. There is an extensive literature comparing methods of estimating the mean and variance of a single left-censored measure, with the MLE generally considered the preferred method. Unfortunately, as the number of measures p increases, MLEs become computationally impractical. Single point imputation methods are commonly used in this case. To remedy this problem, we propose maximum marginal likelihood estimates (MMLEs) for estimating means and variances and maximum pairwise pseudo-likelihood estimates (MPPLEs) for the covariance and correlation matrices. Standard bootstrap confidence intervals are proposed for the correlation parameters. Extensive investigation of the bivariate case shows that the MMLE/MPPLEs exhibit low bias in almost all cases but have some bias under severe left-censoring and small sample sizes. The MMLE/MPPLE routine almost always converges and is fast. In contrast, the MLE routine is slower, has a much lower convergence rate under heavy censoring and poorer properties when positively correlated variables are heavily censored in small samples (Table 2). For $p > 2$ the proposed methods show little bias and are preferable over commonly used single-imputation methods. The coverage of the confidence intervals is close to the nominal level. In summary, the proposed pseudolikelihood-based methodology is a competitor to the MLE in the bivariate case ($p = 2$), is very feasible for $p > 2$, and generally outperforms the imputation methods. Application of the proposed methods to the PCB data set uncovered some very interesting unknown correlational patterns that we hope to investigate in later-year data as they become available.

ACKNOWLEDGEMENTS

The authors wish to thank an associate editor and two referees for their helpful suggestions that improved the content and discussion of ideas in this paper.

REFERENCES

- Ampleman MD, Martinez A, DeWall J, Rawn DFK, Hornbuckle KC, Thorne PS. (2014). Inhalation and dietary exposure to PCBs in urban and rural cohorts via congener-specific measurements. *Environmental Science and Technology*, in press.

- Efron B, Tibshirani R. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.
- EFSA. (2010). Management of left-censored data in dietary exposure assessment of chemical substances. EFSA Journal 8(3):1557 Ninety-six page scientific report of the European Food Safety Authority.
- Genz A (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics* 1, 141–149.
- Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T. (2011). mvtnorm: Multivariate Normal and t Distributions. R package version 0.9–9991. URL <http://CRAN.R-project.org/package=mvtnorm>
- Gong G, Samaniego FJ. (1981). Pseudo maximum likelihood estimation: theory and applications. *Annals of Statistics* 9, 861–869.
- Hald A (1949). Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point, *Skandinavisk Aktuarietidskrift* 32, 119–134.
- Helsel DR. (2005). More than obvious: Better methods for interpreting nondetect data. *Environmental Science and Technology* 39, 419A–423A.
- Hewett P. (2006). Analysis of censored data. Appendix VIII of *A Strategy for Accessing and Managing Occupational Exposures*, (eds.) Bullock WH, Ignacio JS (third edition).
- Hewett P, Ganser GH. (2007). A comparison of several methods for analyzing censored data. *The Annals of Occupational Hygiene* 51, 611–632. [PubMed: 17940277]
- Holden LR, Graham JA, Whitmore RW, Alexander WJ, Pratt RW, Liddle SK, Piper LL. (1992). Results of the National Alachlor Well Water Survey. *Environmental Science and Technology* 26, 935–943.
- Hornung RW, Reed LD. (1990). Estimation of average concentration in the presence of nondetectable values. *Applied Occupational and Environmental Hygiene* 5, 46–51.
- Kerr GD. (1994). Missing dose from mortality studies of radiation effects among workers at Oak Ridge National Laboratory. *Health Physics* 66, 206–208. [PubMed: 8282563]
- Lauby-Secretan B, Loomis D, Grosse Y, El Ghissassi F, Bouvard V, Benbrahim-Tallaa L, Guha N, Baan R, Mattock H, Straif K. (2013). Carcinogenicity of polychlorinated biphenyls and polybrominated biphenyls. *The Lancet Oncology*, 14:4, 287–288. [PubMed: 23499544]
- Lyles RH, Williams JK, Chuachoowong R. (2001). Correlating two viral load assays with known detection limits. *Biometrics* 57, 1238–1244. [PubMed: 11764265]
- Marek RF, Thorne PS, Wang K, DeWall J, Hornbuckle KC. (2013). PCBs and OH-PCBs in serum from children and mothers in urban and rural U.S. communities. *Environmental Science & Technology* 47(7), 3353–3361. Erratum in: *Environmental Science & Technology*, 2013 Aug 20; 47(16), 9555–6. [PubMed: 23452180]
- Nehls GJ, Akland GG. (1973). Procedures for handling aerometric data. *Journal of the Air Pollution and Control Association* 23, 180–184.
- R: A language and environment for statistical computing. R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2005.
- SAS Institute. (2004). *The Lifereg Procedure*. Version 9.1.2. Cary, North Carolina: SAS Institute.
- Shumway RH, Azari RS, Johnson P. (1989). Estimating mean concentrations for environmental data with detection limits. *Technometrics* 31, 347–356.
- Shumway RH, Azari RS, Kayhanian M. (2002). Statistical approaches to estimating mean water quality concentrations with detection limits. *Environmental Science and Technology* 36, 3345–3353. [PubMed: 12188364]
- Thorne PS, Perry SS, Saito R, O'Shaughnessy PT, Mehaffy J, Metwali N, Keefe T, Donham KJ, Reynolds SJ. (2010). Evaluation of the Limulus amoebocyte lysate and recombinant factor C assays for assessment of airborne endotoxin. *Applied and Environmental Microbiology*, 76(15), 4988–4995. [PubMed: 20525858]

Figure 1. PPMLEs of the correlations among all congener groupings. Left panel contains the correlations of groupings A - G with all 16 congeners. Right panel contains the correlations of groupings H - P with all congeners.

Table 1

Average parameter estimates for a single bivariate normal; nlmcode = 1 or 2.

n	ρ	%LC [†]	MMLE/MPPLE					MLE [*]				
			μ_1	μ_2	σ_1^2	σ_2^2	ρ	μ_1	μ_2	σ_1^2	σ_2^2	ρ
30	-0.4	25,25	-0.003	0.007	1.972	2.975	-0.400	-0.003	0.008	1.970	2.972	-0.401
		50,50	-0.020	-0.026	2.012	3.011	-0.393	-0.018	-0.023	2.007	3.000	-0.395
		75,75	-0.026	-0.032	2.095	3.168	-0.541	-0.001	0.024	2.038	2.988	-0.511
		25,75	-0.006	-0.003	1.972	3.084	-0.406	-0.005	0.001	1.964	3.077	-0.408
	0.8	25,25	-0.010	-0.014	2.025	3.004	0.796	-0.010	-0.012	2.030	3.002	0.797
		50,50	-0.013	-0.017	2.056	3.031	0.794	-0.016	-0.020	2.062	3.040	0.797
		75,75	-0.026	-0.036	2.089	3.096	0.779	-0.029	-0.055	2.126	3.210	0.784
		25,75	-0.012	-0.009	1.985	3.062	0.794	-0.011	-0.031	1.988	3.157	0.803
100	-0.4	25,25	0.003	-0.005	2.002	2.974	-0.392	0.003	-0.004	2.002	2.971	-0.392
		50,50	-0.005	-0.002	2.010	2.982	-0.404	-0.004	-0.001	2.008	2.978	-0.405
		75,75	-0.009	-0.023	2.053	3.072	-0.439	-0.005	-0.015	2.050	3.050	-0.433
		25,75	-0.005	-0.007	1.990	3.073	-0.396	-0.005	-0.007	1.989	3.071	-0.397
	0.8	25,25	-0.007	-0.010	2.001	3.007	0.799	-0.007	-0.011	2.001	3.006	0.800
		50,50	-0.008	0.000	2.011	3.018	0.799	-0.008	-0.001	2.011	3.019	0.800
		75,75	-0.013	-0.013	2.009	3.011	0.791	-0.014	-0.014	2.010	3.012	0.793
		25,75	0.002	-0.009	1.991	3.054	0.801	0.002	-0.012	1.990	3.065	0.803

* Starting values for MLE involved $L/2$ imputation of left-censored values[†] %LC: Percents that Y_1, Y_2 were left censored

Table 2

Root mean squared errors of the estimates for a single bivariate normal; nlmcode=1,2

n	ρ	%LC [†]	MMLE/MPPLE					MLE [*]				
			μ_1	μ_2	σ_1^2	σ_2^2	ρ	μ_1	μ_2	σ_1^2	σ_2^2	ρ
30	-0.4	25,25	0.281	0.336	0.666	0.974	0.175	0.280	0.337	0.663	0.975	0.175
		50,50	0.344	0.404	0.884	1.321	0.201	0.341	0.404	0.885	1.315	0.201
		75,75	0.598	0.814	1.421	2.736	0.409	0.566	0.787	1.360	2.708	0.403
		25,75	0.278	0.709	0.681	2.151	0.230	0.278	0.692	0.677	2.088	0.230
	0.8	25,25	0.274	0.340	0.672	0.980	1.198	0.271	0.337	0.659	0.955	1.200
		50,50	0.330	0.416	0.866	1.368	1.198	0.333	0.415	0.864	1.328	1.201
		75,75	0.603	0.744	1.537	2.149	1.190	0.663	0.845	2.048	3.102	1.195
		25,75	0.272	0.692	0.663	2.054	1.199	0.272	0.688	0.661	2.246	1.207
100	-0.4	25,25	0.153	0.179	0.341	0.507	0.090	0.153	0.179	0.341	0.505	0.090
		50,50	0.178	0.218	0.458	0.685	0.106	0.177	0.217	0.455	0.681	0.106
		75,75	0.304	0.370	0.715	1.046	0.212	0.303	0.365	0.714	1.036	0.204
		25,75	0.148	0.371	0.346	1.127	0.117	0.148	0.368	0.347	1.107	0.117
	0.8	25,25	0.150	0.183	0.337	0.535	1.200	0.150	0.183	0.338	0.525	1.200
		50,50	0.178	0.219	0.435	0.662	1.200	0.176	0.217	0.426	0.645	1.201
		75,75	0.313	0.365	0.731	1.055	1.193	0.306	0.358	0.713	1.027	1.195
		25,75	0.147	0.365	0.328	1.081	1.202	0.146	0.349	0.328	1.046	1.205

* Starting values for MLE involved $L/2$ imputation of left-censored values

[†] %LC: Percents that Y_1, Y_2 were left censored

Table 3

Substudy 3 : Effect of starting value and convergence code. 75 Percent of both Y_1 and Y_2 were left censored, $n=30$, $\rho = -0.4$.

Method	Start	nlm.code	% conv	μ_1	μ_2	σ_1^2	σ_2^2	ρ	
Average parameter estimates									
MMLE/MPPLE		1	100.0	-0.026	-0.032	2.095	3.168	-0.541	
MLE	L/2	1	86.9	0.025	0.051	1.974	2.942	-0.486	
MMLE/MPPLE [*]				0.021	0.050	1.985	2.949	-0.480	
MLE	L/2	1-2	90.5	0.006	0.025	2.006	2.990	-0.507	
MMLE/MPPLE [*]				0.002	0.014	2.016	3.038	-0.498	
MLE	L	1	68.6	0.020	0.025	2.057	3.133	-0.366	
MLE	L	1-2	94.3	-0.992	-0.741	18.185	17.577	-0.539	
MLE	MMP [†]	1	99.6	-0.024	-0.032	2.089	3.167	-0.541	
	MMP [†]	1-2	100.0	-0.023	-0.031	2.086	3.162	-0.542	
MMLE/MPPLE		1	100.0	0.598	0.814	1.421	2.736	0.409	
MLE	L/2	1	86.9	0.539	0.745	1.295	2.536	0.392	
MMLE/MPPLE				0.537	0.739	1.300	2.489	0.385	
MLE	L/2	1-2	90.5	0.566	0.762	1.329	2.535	0.403	
MMLE/MPPLE				0.568	0.792	1.332	2.610	0.392	
MLE	L	1	68.6	0.566	0.838	1.366	2.961	0.343	
MLE	L	1-2	94.3	2.896	2.867	155.112	279.545	0.429	
MLE	MMP [†]	1	99.6	0.600	0.819	1.419	2.772	0.409	
MLE	MMP [†]	y	1-2	100.0	0.599	0.819	1.417	2.772	0.409

* The lines marked MMLE/MPPLE* denote the MMLEs/MPPLE based on the same data sets as the MLEs in the line just above them, providing a head-to-head comparison.

[†]When MMP appears in the start column, the MMLEs/MPPLE are used as startuin value for MLE routine.

Table 4

Averages and root mean squared errors for the estimates of the multivariate normal means and variances by method.

Measure	%LC [*]	True	Average $\hat{\mu}$			RMSE $\hat{\mu}$		
			MMLE [†]	$L/2$	$L/\sqrt{2}$	MMLE	$L/2$	$L/\sqrt{2}$
1	25.2	0	-0.008	-0.029	0.058	0.107	0.106	0.107
2	49.8	0	-0.001	0.054	0.226	0.128	0.104	0.238
3	74.9	0	0.003	0.305	0.564	0.214	0.311	0.566
4	24.7	0	0.005	-0.018	0.068	0.106	0.103	0.114
5	50.1	0	-0.005	0.054	0.228	0.126	0.102	0.239
6	75.0	0	-0.006	0.304	0.564	0.213	0.310	0.566

Measure	%LC	True	Average $\hat{\sigma}^2$			RMSE $\hat{\sigma}^2$		
			MMLE	$L/2$	$L/\sqrt{2}$	MMLE	$L/2$	$L/\sqrt{2}$
1	25.2	1	1.003	1.005	0.794	0.181	0.122	0.231
2	49.8	1	0.998	0.737	0.509	0.240	0.283	0.499
3	74.9	1	1.001	0.372	0.227	0.362	0.634	0.776
4	24.7	1	0.995	1.002	0.793	0.185	0.127	0.234
5	50.1	1	1.019	0.745	0.515	0.234	0.274	0.492
6	75.0	1	1.015	0.372	0.227	0.357	0.633	0.775

* %LC : Percent left censoring

† Percent coverage of MMLE-based 95% CI of μ 's: 94.9, 94.7, 93.2, 95.0, 95.6, 93.7

Table 5

Averages and root mean squared errors for estimates of the multivariate normal correlations.

Measures	True	Average $\hat{\rho}$			MPPLE CI Cover*	RMSE $\hat{\rho}$		
		MPPLE	$L/2$	$L/\sqrt{2}$		MPPLE	$L/2$	$L/\sqrt{2}$
1,2	0.7	0.697	0.636	0.644	92.9%	0.064	0.090	0.086
1,3	0.7	0.696	0.545	0.555	92.2%	0.078	0.171	0.164
1,4	0.0	0.004	0.004	0.004	93.7%	0.108	0.101	0.101
1,5	0.0	-0.010	-0.008	-0.007	95.0%	0.113	0.099	0.099
1,6	0.0	0.000	0.002	0.002	95.7%	0.133	0.099	0.100
2,3	0.7	0.697	0.572	0.582	93.6%	0.081	0.153	0.148
2,4	0.0	0.004	0.004	0.005	95.0%	0.114	0.100	0.100
2,5	0.0	-0.009	-0.006	-0.006	94.7%	0.121	0.099	0.099
2,6	0.0	0.000	0.003	0.004	95.8%	0.141	0.099	0.099
3,4	0.0	0.005	0.004	0.004	94.0%	0.136	0.101	0.100
3,5	0.0	-0.009	-0.003	-0.003	94.7%	0.148	0.101	0.101
3,6	0.0	-0.016	-0.001	-0.001	98.3%	0.173	0.096	0.097
4,5	0.4	0.396	0.355	0.356	92.7%	0.102	0.104	0.106
4,6	0.4	0.396	0.305	0.305	94.7%	0.113	0.132	0.135
5,6	0.4	0.394	0.304	0.303	94.8%	0.122	0.141	0.146

* Coverage of MPPLE-based 95% Bootstrap confidence interval of correlation

Percent left censoring by measure: (1) 25.2, (2) 49.8, (3) 74.9, (4) 24.7, (5) 50.1, (6) 75.0

Table 6Analysis of PCB congeners^{*} having at least 25% uncensored values

Peak	PCB Congener	Percent	MMLE		Prin Comp Loadings	
	Grouping [†]	Uncensored	Mean	SD	Comp1	Comp2
A	P11	26.6	1.517	0.400	0.031	−0.328
B	P28+20	43.7	1.394	0.309	0.067	−0.255
C	P22	30.2	0.824	0.214	0.021	−0.189
D	P70+74+ ⁶¹ + ⁷⁶	26.6	1.666	0.585	0.250	−0.592
E	P66	42.7	1.251	0.394	0.100	−0.491
F	P99+83	71.9	1.478	0.360	0.153	−0.131
G	P118	62.5	1.658	0.493	0.210	−0.238
H	P138+ ¹²⁹ + ¹⁶³	80.2	1.745	0.580	0.348	0.044
I	P146	48.4	0.863	0.396	0.209	0.032
J	P153+ ¹⁶⁸	88.5	1.800	0.608	0.370	0.115
K	P156+ ¹⁵⁷	32.8	0.998	0.393	0.233	0.032
L	P170	46.9	0.959	0.544	0.326	0.094
M	P180+ ¹⁹³	77.6	1.422	0.661	0.400	0.210
N	P187	53.1	0.930	0.584	0.342	0.196
O	P199+ ¹⁹⁸	28.6	0.825	0.478	0.268	0.115
P	P203	25.5	0.725	0.393	0.207	0.066
PC Standard Deviation					1.582	0.618
PC Proportion of Variance					0.686	0.105
PC Cumulative Proportion					0.686	0.791

^{*} Cube root of congener values used to achieve normality.

[†] The major congener by Aroclor production is listed first. Minor congeners are superscripted.