Nonparametric maximum likelihood estimation for the multi-sample Wicksell corpuscle problem

BY KWUN CHUEN GARY CHAN

Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.

kcgchan@u.washington.edu

AND JING QIN

Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland 20892, U.S.A.

jingqin@niaid.nih.gov

SUMMARY

We study nonparametric maximum likelihood estimation for the distribution of spherical radii using samples containing a mixture of one-dimensional, two-dimensional biased and three-dimensional unbiased observations. Since direct maximization of the likelihood function is intractable, we propose an expectation-maximization algorithm for implementing the estimator, which handles an indirect measurement problem and a sampling bias problem separately in the E- and M-steps, and circumvents the need to solve an Abel-type integral equation, which creates numerical instability in the one-sample problem. Extensions to ellipsoids are studied and connections to multiplicative censoring are discussed.

Some key words: Abel-type integral equation, expectation-maximization algorithm, indirect measurement, particle size.

1. INTRODUCTION

Wicksell (1925) studied the estimation of the size of spherical corpuscles found in anatomical samples of tissues in organs such as the spleen, thymus and pancreas. The samples were prepared by thin cross-sections of tissues and circular profiles of the corpuscles were observed. Wicksell identified two statistical challenges: an indirect measurement problem and a sampling bias problem, which will be described in detail in §2. He proposed a mathematical formulation of the problem where the distribution of spherical radii of interest and the distribution of two-dimensional observations are related through an Abel-type integral equation. However, the problem is considered to be ill-posed, as a naive estimator defined by plugging in the empirical distribution of observations is often badly behaved, as illustrated in Figure 1a.

The Wicksell corpuscle problem has numerous applications because there are many practical situations that involve indirect measurement of three-dimensional objects. Although recent technological advancements have opened the possibility of direct three-dimensional measurements, indirect measurements from lower dimensions are still widely used for practical reasons. A comparison of radii distribution of particles from two-dimensional and three-dimensional measurements of nickel-based superalloy Inconel 100 was considered in Tucker et al. (2012). Using
different technologies, it is now possible to combine observations from different dimensions to improve estimation accuracy.

The statistical literature on one-sample estimation of the corpuscle problem is vast; see Chiu et al. (2013) for a comprehensive review. It mainly considers one-sample estimation with two-dimensional observations, and many statistical and numerical procedures have been developed to overcome the ill-posed nature of the problem. For example, Hall and Smith (1988), Van Es and Hoogendoorn (1990) and Golubev and Levit (1998) considered kernel smoothing methods, Nychka et al. (1984) studied a spline-based method, Antondiadis, Fan and Gijbels (2001) proposed a wavelet method, and Groeneboom and Jongbloed (1995) considered an isotonized estimator. Many other statistical and numerical methods are surveyed in Chiu et al. (2013), who comment that no method has clear advantages. While each method has distinctive merits in one-sample estimation, those methods typically regularize estimators obtained from an Abel-type integral equation, and such a representation is not readily extended to multiple samples.

We study maximum likelihood estimation in the multi-sample problem, where observations are collected from a combination of one-dimensional, two-dimensional and three-dimensional observations. Direct maximization of the likelihood function is intractable, and we will present an EM algorithm to combine three-sample observations, with special cases yielding new two-sample and one-sample estimators. In one-sample Wicksell problems, Vardi et al. (1985), Wilson (1989) and Silverman et al. (1990) developed EM-type algorithms, but their approaches differ from the proposed method in key aspects. First, all three papers considered a categorical data model, discretizing the domain of interest into finite number of bins and applying the EM algorithm for categorical data (Little and Rubin, 2002, Ch. 13). To reduce the arbitrariness of binning, they considered smoothing after each iteration. In contrast, our method does not require binning and smoothing, estimating the distribution function directly. Also, our method does not require numerical inversion of discretized Abel-type integral equation. As a by-product, the algorithm also ensures that the resulting nonparametric maximum likelihood estimator is monotone, as shown in Figure 1b.
2. NOTATIONS AND PROBLEM FORMULATION

Suppose that spherical particles of different radii are randomly distributed in \( \mathbb{R}^3 \), where the centers of the spheres are distributed according to a stationary spatial Poisson process. While the Poisson process assumption is imposed in most statistical papers, it can be relaxed, as discussed in \( \S 3 \). We are interested in estimating the distribution function of the radii of the particles, which is assumed to have a finite second moment and a density function \( f_R(R) \). Suppose we sample the spheres using a two-dimensional planar sample, and we observe only the circular profiles of the spheres intersecting the plane. Let \( r \) be the radii of the two-dimensional circular profiles. Two statistical challenges are present. First, the spheres with larger radii are more likely to be sampled, and the spheres are sampled with probability proportional to their radii. Therefore, the sampling distribution of the spherical radii is

\[
f_{R}^{S}(R) = \frac{R f_{R}(R)}{\int_{0}^{\infty} R f_{R}(R) dR}, \quad R > 0.
\]

Second, the radii \( r \) of the two-dimensional circular profiles are indirect measurements and are always smaller than the radii of the spheres being sampled. Given that a sphere with radius \( R \) is being sampled, Wicksell (1925) showed that

\[
f_{S}(r | R) = \frac{r}{R^{2} - r^{2})^{1/2}}, \quad 0 < r \leq R.
\]

Therefore, the sampling density of \( r \) is given by

\[
f_{S}(r) = \int_{r}^{\infty} \frac{r f_{R,2}^{S}(R) dR}{R^{2} - r^{2})^{1/2}} = \int_{0}^{\infty} R f_{R}(R) dR \int_{r}^{\infty} \frac{f_{R}(R) dR}{(R^{2} - r^{2})^{1/2}}.
\]  \( \quad \text{(1)} \)

Suppose, in addition, that we sample using a one-dimensional linear probe. Let \( l \) be the half length of the traces where the linear probe intersects the spheres. The spheres are now sampled proportional to their squared radii,

\[
f_{S}^{R}(R) = \frac{R^{2} f_{R}(R)}{\int_{0}^{\infty} R^{2} f_{R}(R) dR}, \quad R > 0,
\]

and the half-lengths of the measurements are again always smaller than the radii of the spheres. Given that a sphere with radius \( R \) is sampled, Watson (1971) showed that the sampling distribution of \( l \) given \( R \) is

\[
f_{S}(l | R) = \frac{2l}{R^{2}}, \quad 0 < l \leq R,
\]

so the sampling distribution of \( l \) is

\[
f_{S}(l) = \frac{2l \{1 - F_{R}(l)\}}{\int_{0}^{\infty} R^{2} f_{R}(R) dR}, \quad l > 0.
\]

We are interested in estimating \( F_{R} \), the population distribution of \( R \). Suppose we have an unbiased three-dimensional random sample of independent observations \( R_{i} \) \((i = 1, \ldots, n_{3})\), an independent two-dimensional cross-sectional sample \( r_{j} \) \((j = 1, \ldots, n_{2})\), and a one-dimensional linear probe sample \( l_{k} \) \((k = 1, \ldots, n_{1})\). The total sample size is \( n = n_{1} + n_{2} + n_{3} \). The likeli-
hood function is
\[
L = \left\{ \prod_{i=1}^{n_3} f_{R}(R_i) \right\} \left\{ \prod_{j=1}^{n_2} \int_{0}^{\infty} dR \frac{r_j}{R f_{R}(R)} \int_{r_j}^{\infty} \frac{1}{(R^2 - r_j^2)^{1/2}} f_{R}(R) dR \right\} \left\{ \prod_{k=1}^{n_1} \frac{2l_k(1 - F_{R}(l_k))}{\int_{0}^{\infty} R^2 f_{R}(R) dR} \right\}.
\]

While direct maximization of (2) appears to be intractable, we capitalize on a subtle difference between the indirect measurement and the sampling bias problems to derive an EM algorithm for nonparametric maximum likelihood estimation.

3. Maximum likelihood estimation and EM algorithm

To shed light on how to perform maximum likelihood estimation for this problem, we first suppose we can observe the radii \( R_j \) (\( j = 1, \ldots, n_2 \)) of the spheres being sampled by the two-dimensional planar probe, and the radii \( R_k \) (\( k = 1, \ldots, n_1 \)) of the spheres being sampled by the one-dimensional linear probe. Based on the hypothetical complete observations of radii \( R \) for the three samples, the complete data likelihood,
\[
L_C = \left\{ \prod_{i=1}^{n_3} f_{R}(R_i) \right\} \left\{ \prod_{j=1}^{n_2} \frac{R_j f_{R}(R_i)}{R f_{R}(R)} \int_{0}^{\infty} dR \right\} \left\{ \prod_{k=1}^{n_1} \frac{R_k^2 f_{R}(R_i)}{R^2 f_{R}(R)} \int_{0}^{\infty} dR \right\},
\]
is a biased-sample likelihood function, as in Vardi (1985), who also showed that the nonparametric maximum likelihood estimator of \( F_{R}(R) \) based on (3) assigns point mass only to distinct data points of the combined samples. Let \( t_1 < \cdots < t_H \) be the observed distinct data points, \( p_h = dF_{R}(t_h) \) (\( h = 1, \ldots, H \)), and let \( \xi_h, \eta_h \) and \( \nu_h \) denote the multiplicities of the three-dimensional sample, two-dimensional sample and one-dimensional sample at \( t_h \) respectively, i.e., \( \xi_h = \sum_{i=1}^{n_3} I(R_i = t_h), \eta_h = \sum_{j=1}^{n_2} I(R_j = t_h) \) and \( \nu_h = \sum_{k=1}^{n_1} I(R_k = t_h) \). Then the nonparametric maximum likelihood estimator based on the complete \( R_i, R_j \) and \( R_k \) can be found by maximizing
\[
\left( \prod_{h=1}^{H} p_h^{\xi_h + \eta_h + \nu_h} \right) \left( \sum_{h=1}^{H} p_h t_h \right)^{-n_2} \left( \sum_{h=1}^{H} p_h t_h^2 \right)^{-n_1}
\]
subject to the constraints
\[
\sum_{h=1}^{H} p_h = 1, \quad p_h > 0, \quad h = 1, \ldots, H.
\]

Therefore, if the radii \( R \) are observed in all samples, the estimation for \( F_{R} \) reduces to that in Vardi (1985) for biased sampling problems, for which a computationally efficient algorithm proposed by Mallows (1985) was shown to converge to the nonparametric maximum likelihood estimate by Davidov and Illiopoulos (2010). However, we cannot simply proceed by using the above method for the multi-sample corpuscle problem, since \( \eta_h \) and \( \nu_h \) are unknown due to the indirect measurement problem. Noting that
\[
\Pr(R_j = t_h \mid r_j) = \frac{(t_h^2 - r_j^2)^{-1/2} I(t_h > r_j)p_h}{\sum_{h'=1}^{H}(t_{h'}^2 - r_j^2)^{-1/2} I(t_{h'} > r_j)p_{h'}},
\]
\[
\Pr(R_k = t_h \mid l_k) = \frac{I(t_h > l_k)p_h}{\sum_{h'=1}^{H} I(t_{h'} > l_k)p'_{h}}, \quad h = 1, \ldots, H,
\]
the log-likelihood based on (4) is linear in $\eta_h$ and $\nu_h$, we can proceed with an EM algorithm:

\textbf{Algorithm 1.} EM algorithm for computing the spherical radii distribution.

\textbf{Step 1.} Initialize $p_h^{(0)} = 1/H$ ($h = 1, \ldots, H$).

\textbf{Step 2.} For $m \geq 1$,

\begin{itemize}
  \item[(E-step)] Calculate
  \begin{align*}
  \eta_h^{(m)} &= \sum_{j=1}^{n_2} \frac{p_h^{(m-1)}(t_h^2 - t_j^2)^{-1/2}I(t_h > r_j)}{\sum_{h'=1}^{H} p_{h'}^{(m-1)}(t_{h'}^2 - r_j^2)^{-1/2}I(t_{h'} > r_j)}, \\
  \nu_h^{(m)} &= \sum_{k=1}^{n_1} \frac{I(t_h > l_k)p_h^{(m-1)}}{\sum_{h'=1}^{H} I(t_{h'} > l_k)p_{h'}^{(m-1)}}.
  \end{align*}

\end{itemize}

\textbf{Step 3.} Repeat Step 2 until a convergence criterion is met. We denote the final estimate by $\hat{p}_h$ ($h = 1, \ldots, H$). The nonparametric maximum likelihood estimate of $F_R(z)$ is

$$
\hat{F}_R(z) = \sum_{h=1}^{H} \hat{p}_h I(t_h \leq z).
$$

In the Appendix, we show that the log-likelihood function is strictly concave, and thus has a unique maximizer $\hat{p}$. The estimate based on the EM algorithm converges to the nonparametric maximum likelihood estimate $\hat{p}$, following Csiszár and Tusnády (1984), since the set of all probability measures over which the likelihood is maximized is convex.

While the Poisson assumption was employed in the original derivation of Wicksell (1925) and has been assumed in most subsequent statistical papers, the core of Wicksell’s solution, the Abel-type integral equation (1) which represents the combination of both the measurement and the sampling bias problems, can be developed in more general settings. Mecke and Stoyan (1980) showed that (1) can be derived when the centers of the spheres follow stationary point processes. This allows the case where overlapping spheres are removed (Bartlett, 1974) and the remaining spheres are weakly correlated. Jensen (1984) showed that (1) can be derived when the non-overlapping spheres are assumed to be deterministic, but the location of the planar probe is random. Under these assumptions, the measurement and the sampling bias problems remain the same as in Wicksell’s original problem, so the proposed method still maximizes the likelihood.
function (2), which is an independence likelihood (Lindsay, 1988) when the spheres are correlated. In one-sample problems, Heinrich (2007) showed that certain maximum independence likelihood estimators are asymptotically normal when the spheres are weakly correlated. Simulations were conducted to evaluate the proposed method when the Poisson assumption is violated; see §6.1.

Remark 1. Spherical assumptions are common in stereology. The sensitivity of the deviation from the random spheres approximation was examined in Anderssen and Jakeman (1974), who concluded that the approximation is quite reliable. For particles with random shape, there is no general relationship relating to the size distribution of three-dimensional particles and two-dimension sections. Our method can be generalized to some particular cases, such as ellipsoids, where an explicit relationship between the sizes of three-dimensional particles and two-dimensional sections is available.

4. EXTENSION TO ELLIPSOIDS

Although widely studied, the original Wicksell problem only considers univariate size distributions with a constant spherical shape. For particles with variation in both shape and size, it is desirable to estimate the joint distribution of a size and a shape measure. A useful model is the ellipsoid model, which has been studied since Wicksell (1926) under independence between size and shape, while Cruz-Orive (1976) studied a general mathematical formulation. Although the problem is much more complicated than the spherical case, two-dimensional indirect observations of three-dimensional ellipsoids are subject to the same statistical problems, so the proposed EM algorithm for spheres can be extended to ellipsoids. To illustrate the ideas, we first consider the case where we only have two-dimensional indirect observations. As discussed in Cruz-Orive (1976), one cannot nonparametrically identify the joint distribution of axes of a triaxial ellipsoid. Instead, one can only identify the joint distribution of the major semiaxes \( a \) and the minor semiaxes \( b \) of a biaxial ellipsoid, which could be prolate or oblate. For simplicity we consider prolate ellipsoids; the derivation is very similar for oblate ellipsoids. We adopt the reparameterization of Cruz-Orive (1976) and consider estimation of the joint distribution of \( b \) and the eccentricity parameter \( x^2 = 1 - (b/a)^2 \). The joint density and distribution of \((b, x^2)\) are denoted by \( f_{b,x^2}(b, x^2) \) and \( F_{b,x^2}(b, x^2) \). The two-dimensional observed ellipses are \((m_i, y_i^2)\) \(i = 1, \ldots, n_2\).

Let \( \phi(b, x^2) = b(1 - x^2)^{-1/2} \int_0^{\pi/2} (1 - x^2 \sin^2 \theta)^{1/2} \sin \theta \, d\theta \). If we can observe the ellipsoids that intersect the probe, Cruz-Orive (1976) showed that the sampling distribution of \((b, x^2)\) is

\[
    f_{b,x^2,2}^S(b, x^2) = \frac{\phi(b, x^2) f_{b,x^2}(b, x^2)}{\int_0^\infty \int_0^1 \phi(b, x^2) f_{b,x^2}(b, x^2) \, dx^2 \, db}.
\]

(8)

Given an ellipsoid with minor semiaxis \( b \) and eccentricity \( x^2 \) that intersects the two-dimensional probe, the observed \((m, y^2)\) is subject to indirect measurement such that \( 0 \leq m < b \) and \( 0 \leq y^2 < x^2 \), and

\[
    f^S(m, y^2 \mid b, x^2) = \frac{1}{2\phi(b, x^2)} \left\{ \frac{m}{(b^2 - m^2)^{1/2}} \right\} \left[ \frac{(1 - y^2)^{1/2}}{x \{(1 - x^2)(x^2 - y^2)^{1/2}\}} \right].
\]

Therefore, the sampling density of \((m, y^2)\) is

\[
    f_{m,y^2}^S(m, y^2) = \frac{1}{2\phi} \int_m^\infty \int_{y^2}^1 \left\{ \frac{m}{(b^2 - m^2)^{1/2}} \right\} \left[ \frac{(1 - y^2)^{1/2}}{x \{(1 - x^2)(x^2 - y^2)^{1/2}\}} \right] f_{b,x^2}(b, x^2) \, dx^2 \, db,
\]
Multi-sample corpuscle problem

where $\phi = \int_{0}^{\infty} \int_{0}^{1} \phi(b, x^2) f_{b,x^2}(b, x^2) db dx^2$.

In an EM-algorithm, the E-step can be constructed by considering the sampling conditional density of $(b, x^2)$ given the observed $(m, y^2)$,

$$f^S(b, x^2 | m, y^2) = \frac{x^{-1} \{(b^2 - m^2)(1 - x^2)(x^2 - y^2)\}^{-1/2} f_{b,x^2}(b, x^2)}{\int_{m}^{\infty} \int_{y_i}^{1} x^{-1} \{(b^2 - m^2)(1 - x^2)(x^2 - y^2)\}^{-1/2} f_{b,x^2}(b, x^2) db dx^2} \quad (9)$$

and the M-step can be constructed by (8). To maximize the likelihood function over the observed data points, let $p_i = F_{b,x^2}(db_i, dx_i^2), i = 1, \ldots, n_2$. The EM algorithm is stated as follows.

Algorithm 2. EM algorithm for ellipsoids.

Step 1. Initialize $p_i^{(0)} = 1/n_2 (i = 1, \ldots, n_2)$.

Step 2. For $k \geq 1$,

(E-step) Following (9), the E-step is

$$\eta_i^{(k)} = \sum_{j=1}^{n_2} \frac{p_i^{(k-1)} m_i^{-1} \{(m_i^2 - m_j^2)(1 - y_i^2)(y_i^2 - y_j^2)\}^{-1/2} I(m_i > m_j, y_i > y_j)}{\sum_{j'=1}^{n_2} p_{j'}^{(k-1)} m_{j'}^{-1} \{(m_{j'}^2 - m_j^2)(1 - y_{j'}^2)(y_{j'}^2 - y_j^2)\}^{-1/2} I(m_{j'} > m_j, y_{j'} > y_j)} \quad . \quad (10)$$

(M-step) Following (8), the M-step is

$$p_i^{(k)} = \frac{\eta_i^{(k)} \phi^{-1}(m_i, y_i^2)}{\sum_{j=1}^{n_2} \eta_j^{(k)} \phi^{-1}(m_j, y_j^2)} \quad .$$

Step 3. Repeat Step 2 until a convergence criterion is met. We denote the final estimate by $\hat{p}_1, \ldots, \hat{p}_{n_2}$. The corresponding estimate of $F_{b,x^2}(w, z)$ is

$$\hat{F}_{b,x^2}(w, z) = \sum_{i=1}^{n_2} \hat{p}_i I(m_i \leq w, y_i^2 \leq z) \quad .$$

Remark 2. We are often interested in estimating the distribution function of certain summary variables, such as axial ratio or volume. In general, these are functions of $b$ and $x^2$. Since the joint distribution of $(b, x^2)$ can be estimated, the distribution of functions of $(b, x^2)$ can also be estimated. For example, let $H$ be the ratio of the minor axis to the major axis. The distribution function of $H$ can be estimated by

$$\hat{F}_H(h) = \int_{w=0}^{\infty} \int_{z=1-h^2}^{1} \hat{F}_{b,x^2}(dw, dz) \quad , \quad 0 < h \leq 1 \quad .$$

For another example, the distribution function of $V$, the volume of an ellipse, can be estimated by

$$\hat{F}_V(v) = \int_{z=0}^{1} \int_{w=0}^{h(v, z)} \hat{F}_{b,x^2}(dw, dz) \quad , \quad v > 0 \quad ,$$

where $h(v, z) = \{3/(4\pi)\}^{1/3}(1 - z)^{1/6}v^{1/3}, v > 0$ and $0 \leq z \leq 1$. McGarry et al. (2014) considered the estimation of univariate summary measures that are functions of radii and heights of cylinders. Their method is specially designed for the case where the height of the cylinders does not suffer from an indirect measurement problem, and cannot be directly extended to estimate the axial ratio distribution or the volume distribution of ellipsoids where a bivariate measurement problem is present.
Remark 3. As shown in Cruz-Orive (1976), a two-dimensional sample cannot identify the joint trivariate distribution of a principal semiaxes and two principal eccentricities from a sample which is a mixture of the prolate and oblate spheroids. The difficulty is primarily due to the non-identifiability of mixture proportions. Using an unbiased three-dimensional sample, however, allows us to identify the proportion of prolate and oblate spheroids in the samples. When the estimated proportion is treated as fixed, we can modify the E-step to distribute a fraction of masses to prolate and oblate spheroids respectively, and the M-step to reweight a mixture of biased samples.

5. CONNECTIONS TO MULTIPLICATIVE CENSORING

Vardi (1989) and Vardi and Zhang (1992) considered the following multiplicative censoring problem. They assumed that two independent samples are available: $X_1, \ldots, X_m$ are independent and identically distributed complete uncensored observations with density $g_R(x)$, and $Y_1 = U_1Z_1, \ldots, Y_n = U_nZ_n$ are incomplete observations where $U_1, \ldots, U_n$ are independent standard uniform distributed and independent of $Z_1, \ldots, Z_n$, which have density $g$. The incomplete observations are random fractions of the complete observations. Vardi (1989) showed that this multiplicative censoring structure is present in three unrelated applications: nonparametric estimation in renewal processes, deconvolution, and estimation of a monotonic decreasing density. The likelihood function for this problem is

$$\prod_{i=1}^{n} g(x_i) \prod_{j=1}^{m} \int_{y_i}^{\infty} \frac{g(z)}{z} \, dz.$$ 

Since given $Z = z$, $Y$ is $U(0, z)$ distributed, an EM-algorithm was proposed in Vardi (1989) with an E-step assigning weights according to the conditional density of $Z = z$ given $Y = y$, which is $z^{-1}g(z)/\int_y^{\infty} z^{-1}g(z)\, dz$, for $z > y$. 

The multiple corpuscle problem can be viewed as the following multiplicative censoring problem, with independent three-dimensional observations $R_i$ ($i = 1, \ldots, n_3$), two-dimensional observations $r_j = V_j R'_j$ ($j = 1, \ldots, n_2$) and one-dimensional observations $l_k = W_k R_k$ ($k = 1, \ldots, n_1$), where $V$ and $(1 - U^2)^{1/2}$ are equally distributed, $W$ and $U^{1/2}$ are equally distributed, $U$ is uniform distributed, the density function of $R'$ is $f_{R,2}^S$ and the density function of $\tilde{R}$ is $f_{R,1}^S$. It can be shown that

$$\Pr(r < t) = \Pr(R' < t) + \Pr\{R' > t, (1 - U^2)^{1/2} < t/R'\}$$

$$= \int_0^t x f_R(x)\, dx + \int_t^{\infty} [1 - (t/x)^2]^{1/2} x f_R(x)\, dx \int_0^\infty x f_R(x)\, dx,$$

and hence $f_r(t) = t\{\int_0^\infty x f_R(x)\, dx\}^{-1} \int_t^{\infty} (x^2 - t^2)^{-1/2} f_R(x)\, dx$. Also,

$$\Pr(l < t) = \Pr(\tilde{R} < t) + \Pr\{\tilde{R} > t, U < (t/\tilde{R})^2\}$$

$$= \int_0^t x^2 f_R(x)\, dx + t^2 \{1 - F_R(t)\} \int_0^\infty x^2 f_R(x)\, dx,$$

and therefore $f_l(t) = \{\int_0^\infty x^2 f_R(x)\, dx\}^{-1} 2t \{1 - F_R(t)\}$. Thus, the likelihood function of this multiplicative censoring problem is equivalent to (2).

The connection to the multiplicative censoring problem of Vardi (1989) explains the similarities as well as the differences between the proposed EM algorithm and that of Vardi (1989).
In comparison to Vardi’s E-step, which assigns weights to the censored observations according to the conditional density of $Z$ given $Y$, the proposed algorithm assigns weights to the two-dimensional observations based on the conditional density of $R'$ given $r$, and to the one-dimensional observations based on the conditional density of $\tilde{R}$ given $l$. The multiplicative censoring formulation also explains the difference in the M-steps, since in Vardi (1989) $Z$ and $X$ are identically distributed, but $R'$ and $\tilde{R}$ are biased versions of $R$.

6. Numerical Examples

6-1. Simulations

We conducted numerical studies to examine the finite sample properties of the estimators proposed in §3. For each simulation scenario, 5000 independent data sets were generated. In the first set of simulations, each data set consists of independent observations from three-dimensional, two-dimensional and one-dimensional samples. We present the results when $R$ was generated from a uniform distribution. We performed additional simulations for beta-distributed $R$; the results were qualitatively similar and are not presented. Table 1 showed the performance of the proposed three-sample estimator for different sample sizes. In general, the proposed estimator had a negligible small-sample bias at a wide range of percentile points along the distribution of $R$. Also, the sampling bias decreased with sample sizes, supporting the consistency of the estimator. The sampling variability of the estimator decreased with an increase in sample size.

We considered unequal sample sizes among the samples in case (e), which had the same total number of observations as in case (b), but had the same total number of three-dimensional observations as in case (a). Since three-dimensional observations were most informative about the three-dimensional radii distribution, the sample variability of (e) lay between (a) and (b). Comparing (a) and (e), the additional two-dimensional and one-dimensional observations were more informative to the upper tail of the distribution because larger objects are more likely to be sampled due to sampling bias.

Table 2 compares estimators using observations from different samples. We compared the three-sample estimator given in §3, with two-sample and one-sample estimators which are special cases of the proposed method, with the empirical distribution function using only three-dimensional observations. Compared to the estimator using observations from three-dimensional data only, inclusion of additional samples improved estimation efficiencies while the small sam-
Table 2. Pointwise performance of the estimators using observations from different samples.

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Samples</th>
<th>(1,2,3)</th>
<th>(2,3)</th>
<th>(3)</th>
<th>(1,2)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>SD</td>
<td>Bias</td>
<td>SD</td>
<td>Bias</td>
<td>SD</td>
</tr>
<tr>
<td>10</td>
<td>-1</td>
<td>20</td>
<td>1</td>
<td>21</td>
<td>-1</td>
<td>21</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>28</td>
<td>&lt;1</td>
<td>31</td>
<td>-2</td>
<td>34</td>
</tr>
<tr>
<td>50</td>
<td>-1</td>
<td>30</td>
<td>&lt;1</td>
<td>32</td>
<td>-2</td>
<td>37</td>
</tr>
<tr>
<td>70</td>
<td>-1</td>
<td>25</td>
<td>1</td>
<td>26</td>
<td>-1</td>
<td>32</td>
</tr>
<tr>
<td>90</td>
<td>-1</td>
<td>16</td>
<td>-1</td>
<td>17</td>
<td>&lt;1</td>
<td>22</td>
</tr>
<tr>
<td>95</td>
<td>&lt;1</td>
<td>12</td>
<td>-1</td>
<td>13</td>
<td>-1</td>
<td>15</td>
</tr>
<tr>
<td>97.5</td>
<td>-1</td>
<td>9</td>
<td>-1</td>
<td>10</td>
<td>&lt;1</td>
<td>11</td>
</tr>
</tbody>
</table>

The sample sizes were $n_1 = n_2 = n_3 = 200$. The values of Bias and SD were multiplied by 1000, and SD represents the sampling standard deviation. Numbers in the brackets indicate the dimensions of observations being included.

Table 3. Pointwise performance of the estimators when the Poisson process assumption is violated.

<table>
<thead>
<tr>
<th>Percentile</th>
<th>$p = 0$</th>
<th>$p = 0.1$</th>
<th>$p = 0.2$</th>
<th>$p = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>SD</td>
<td>Bias</td>
<td>SD</td>
</tr>
<tr>
<td>10</td>
<td>-1</td>
<td>42</td>
<td>-2</td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>43</td>
<td>-4</td>
<td>42</td>
</tr>
<tr>
<td>50</td>
<td>-1</td>
<td>42</td>
<td>-9</td>
<td>42</td>
</tr>
<tr>
<td>70</td>
<td>-1</td>
<td>36</td>
<td>-10</td>
<td>36</td>
</tr>
<tr>
<td>90</td>
<td>-1</td>
<td>23</td>
<td>-5</td>
<td>23</td>
</tr>
</tbody>
</table>

The sample sizes were $n_2 = n_3 = 100$. The values of Bias and SD were multiplied by 1000, SD represents the sampling standard deviation, and $p$ is the fraction of spheres that are removed due to overlapping.

Next, we study the performance of the estimators when the Poisson process assumption is violated. To generate a weakly correlated population of spheres, we follow a model of Bartlett (1974), where spheres are sequentially generated and a sphere is removed if it overlaps with any existing spheres. The resulting process is a stationary marked point process and the spheres are weakly dependent. We study the performance of the estimators by varying the truncation fraction, that is, the fraction of spheres that are removed. Table 3 shows that estimation bias increases slightly but the sampling variability remains similar when the degree of overlapping increases.

We conducted further simulations to evaluate the rate of convergence of the estimators under the Poisson process assumption. We compared two scenarios: when only two-dimensional observations are available, and when 10% of the observations are three-dimensional and 90% are two-dimensional. According to Groeneboom and Jongbloed (1995), the minimax rate of convergence for two-dimensional observations is $n^{1/2} \log(n)^{-1/2}$, so $n^{1/2} \log(n)^{-1/2}$ times the sampling standard deviation $\sigma_r$ of the estimates $\hat{F}_R(r)$ should stabilize as $n$ increases. That is, $\log(n^{1/2} \sigma_r) \approx k_r + 0.5 \times \log \log n$ for some constant $k_r$. We reran the simulations with $n = 1000, 2000, 3000$ and 5000 and evaluated the sampling standard deviation of the estimated distribution function at the 10, 30, 50, 70 and 90 percentiles of the true distribution. We fitted a linear model with outcome $\log(n^{1/2} \sigma_r)$, predictor $\log \log n$ and dummy variables for different
percentiles. The estimated regression coefficient of log log $n$ from the simulations was 0.62, with 95% confidence interval $(0.35, 0.89)$, consistent with the theoretical predictions. For a combination of three-dimensional and two-dimensional observations, we conjecture that the rate of convergence is $n^{1/2}$, which corresponds to a true regression coefficient of 0. Using 10% three-dimensional and 90% two-dimensional data, the estimated regression coefficient of log log $n$ from the simulations was 0.045, with 95% confidence interval $(-0.07, 0.16)$, so we cannot reject the null hypothesis that the rate of convergence is $n^{1/2}$. We found that the rates of convergence were different in the two scenarios even when the second sample was dominated by two-dimensional observations.

### 6.2. Data analysis

We applied the proposed method to estimate the diameter distribution of a nickel-based superalloy Inconel 100, using combined three-dimensional, two-dimensional and one-dimensional data derived from Tucker et al. (2012). The two-dimensional and one-dimensional observations were obtained by planar and linear sections of the sample materials, and the three samples contain non-overlapping particles. The data contain 84 particles from four one-dimensional sections and 254 particles from a two-dimensional section. Since three-dimensional measurements are more costly to obtain, we included a sample of 120 particles from the three-dimensional observations for illustration.

The estimates of the cumulative distribution are shown in Figure 2. Figure 2(a) shows the estimates for the full support of the diameter distribution. An empirical distribution from the two-dimensional sample overestimated the proportion of particles with small diameter, but not the upper tail of the diameter distribution. Figure 2(b) zooms into the lower tail of the distribution, where the empirical distribution of a two-dimensional sample is biased. The nonparametric maximum likelihood estimate based on the combined samples was nearly identical to the empirical distribution based on the three-dimensional observations for diameter less than 1.5 $\mu m$ and cumulative probability less than 10%. This pattern is very similar to that seen in Table 2.

---

*Multi-sample corpuscle problem*
As shown in Table 2, the efficiency gain of the combined sample estimator is more noticeable in the percentile range greater than 10%, and we observe that the estimated cumulative distribution deviates slightly compared to the empirical distribution based on the three-dimensional observations. Based on 1000 bootstrap replicates, the estimated standard errors for estimating the proportion of particles with diameter less than 2.5 \( \mu \text{m} \) were 0.035 and 0.044 for the combined sample estimate and the three-dimensional sample estimate respectively.

7. Concluding Remarks

The corpuscle problem also arises for objects other than three dimensions. The mathematical formulation of the \( d \)-dimensional corpuscle problem can be found in Heinrich (2007). For any \( d \geq 3 \), the relationships between \( d \), \( d - 1 \) and \( d - 2 \) dimensional observations remain the same. Therefore, our method can be applied to the general \( d \)-dimensional problem with a combination of \( d \), \( d - 1 \) and \( d - 2 \) dimensional observations. Nonetheless, \( d = 3 \) contains most applications of practical interest.

We have shown that the proposed EM algorithm converges to the nonparametric maximum likelihood estimate. We will study its large sample-properties in future research: the single-sample estimators based on three-dimensional, two-dimensional and one-dimensional observations have rates of convergence \( n^{1/2} \), \( n^{1/2}(\log n)^{-1/2} \) and \( n^{1/3} \) respectively, as given in an unpublished 1999 Vrije Universiteit lecture notes by G. Jongbloed. Vardi and Zhang (1992) studied the large-sample properties of a nonparametric maximum likelihood estimator proposed in Vardi (1989) for the two-sample multiplicative censoring problem discussed in §5, in which one-sample observations have different convergence rates of \( n^{1/2} \) and \( n^{1/3} \). Their two-sample estimator results in a \( n^{1/2} \) convergence rate and is asymptotically more efficient than the one-sample estimator with \( n^{1/2} \) rate of convergence. Our setting is substantially different, since their problem did not involve multi-sample sampling bias. As a result, the M-step of Vardi’s algorithm only involves an empirical distribution, whereas the M-step of our algorithm involves nonparametric estimation under multiple biased samples. We expect that the proof of theoretical results for the proposed method will be more difficult than in Vardi and Zhang (1992), but based on their theoretical results and our simulation results, we conjecture that our three-sample estimator has a \( n^{1/2} \) rate of convergence when the sampling fraction \( n_3/n \) converges to a non-negligible proportion \( c > 0 \).

Acknowledgement

The authors thank the editor, an associate editor and three reviewers for their helpful comments and suggestions. They also thank Joe Tucker and Anthony Rollett for providing the Inconel 100 data. The first author is partially supported by the U.S. National Institutes of Health.

Appendix

Concavity of the log-likelihood function

Let \( dH_1(t) = tdF(t)/\int udF(u) \), \( q_h = dH_1(t_h) \) \( (h = 1, \ldots, H) \), and \( \xi_h = \sum_{i=1}^{n_3} I(R_i = t_h) \), \( \zeta_h = \sum_{j=1}^{n_2} I(r_j = t_h) \) and \( \phi_h = \sum_{k=1}^{n_1} I(l_k = t_h) \), which are the multiplicities of the three-dimensional sample, two-dimensional sample and one-dimensional sample at \( t_h \) respectively. Furthermore, let \( p_j = q_j/t_j \), \( P_j = \sum_{k \geq j} p_k/(t_k - t_j)^{1/2} \), \( Q_j = \sum_{k \geq j} p_k \) and \( Q = \sum_{h=1}^H t_h^2 p_h \). The likelihood function can
be written as

\[ L \propto \prod_{h=1}^{H} \phi_h \frac{Q_h}{Q}^{\psi_h}, \]

so the log-likelihood can be expressed as \( \ell = \log L = \ell_1 + \ell_2 + \ell_3 + \ell_4 \) where \( \ell_1 = \sum_{h=1}^{H} \xi_h \log p_h \), \( \ell_2 = \sum_{h=1}^{H} \zeta_h \log Q_h \), \( \ell_3 = \sum_{h=1}^{H} \phi_h \log Q_h \), \( \ell_4 = -n_1 \log Q \). Let \( \ell^\dagger = \ell_1 + \ell_2 + \ell_3 \) and \( H^1 \) be the Hessian of \( \ell^\dagger \). For \( \alpha = (\alpha_1, \ldots, \alpha_H)^T \),

\[
\alpha^T H^1 \alpha = - \sum_{i=1}^{H} \frac{\alpha_i}{p_i^2} - \sum_{i=1}^{H} \left\{ \frac{\alpha_{i+1}}{(t_{i+1}^2 - t_i^2)^{1/2}} + \cdots + \frac{\alpha_H}{(t_H^2 - t_1^2)^{1/2}} \right\}^2 \frac{\zeta_i}{p_i^2}
- \sum_{i=1}^{H} (\alpha_i^2 + \cdots + \alpha_H^2) \frac{\phi_i}{Q_i^2},
\]

and since \( \xi_i \geq 0, \zeta_i \geq 0 \) and \( \phi_i \geq 0 \) with \( \xi_1 + \zeta_1 + \phi_1 > 0 \), this quadratic form is strictly negative unless \( \alpha \equiv 0 \). Therefore \( \ell^\dagger \) is strictly concave. To show the concavity of \( \ell_4 \), we consider the transformation \( \theta_j = \log p_j \), and denote \( H_4 \) be the Hessian of \( \ell_4 \). For \( \alpha = (\alpha_1, \ldots, \alpha_H)^T \),

\[
\alpha^T H_4 \alpha = - \frac{n_1}{Q^2} \left\{ \sum_{i=1}^{H} t_i^2 \exp(\theta_i) \right\} \left\{ \sum_{i=1}^{H} t_i^2 \exp(\theta_i) \alpha_i^2 \right\} - \left\{ \sum_{i=1}^{H} t_i^2 \exp(\theta_i) \alpha_i \right\}^2 \leq 0
\]

by the Cauchy–Schwartz inequality. Since \( \ell^\dagger \) is strictly concave and \( \ell_4 \) is concave, \( \ell = \ell^\dagger + \ell_4 \) is strictly concave.

REFERENCES


[Received x x. Revised x x]