Appl. Statist. (2002)
51, *Part* 4, *pp.* 485–492

Minimum distance estimation of the distribution functions of stochastically ordered random variables

Ronald E. Gangnon and William N. King

University of Wisconsin, Madison, USA

[Received June 2001. Revised April 2002]

Summary. Stochastic ordering of distributions can be a natural and minimal restriction in an estimation problem. Such restrictions occur naturally in several settings in medical research. The standard estimator in such settings is the nonparametric maximum likelihood estimator (NPMLE). The NPMLE is known to be biased, and, even when the empirical cumulative distribution functions nearly satisfy the stochastic orderings, the NPMLE and the empirical cumulative distribution functions may differ substantially. In many settings, this can make the NPMLE seem to be an unappealing estimator. As an alternative to the NPMLE, we propose a minimum distance estimator of distribution functions subject to stochastic ordering constraints. Consistency of the minimum distance estimator is proved, and superior performance is demonstrated through a simulation study. We demonstrate the use of the methodology to assess the reproducibility of gradings of nuclear sclerosis from fundus photographs.

Keywords: Cramér–von Mises distance; Empirical distribution function; Kaplan–Meier estimator; Measurement error; Minimum distance estimation; Stochastic ordering constraints

1. Introduction

In some settings, two or more distribution functions can be assumed to satisfy a stochastic ordering constraint. For example, consider the relationship of the stage in patients diagnosed with cancer to their subsequent survival. It would be expected that more advanced disease would be associated with decreased survival at all subsequent time points. In this paper, we consider a second, perhaps less obvious, situation in which a natural stochastic ordering constraint occurs: the assessment of the variability of a grading system. At the Fundus Photograph Reading Center at the University of Wisconsin, trained graders assess the severity of nuclear sclerosis by using an ordered decimalized scale. One intuitive characterization of the reproducibility of the grading system is based on the conditional distribution of a repeat grading given an original grade. For a highly reproducible grading system, we would expect this distribution to be concentrated near the original grade; for less reproducible systems, we would expect this distribution to be more widely spread. In many grading systems such as the nuclear sclerosis severity scale, we would expect higher original grades to be associated with higher repeat grades and lower original grades to be associated with lower repeat grades.

In the absence of an ordering constraint, the natural estimators of the cumulative distribution functions (CDFs) for each population are the empirical CDFs with complete data and the

Address for correspondence: Ronald E. Gangnon, Department of Biostatistics and Medical Informatics, Statistical Data Analysis Center, University of Wisconsin Medical School, 610 North Walnut Street, Madison, WI 53726, USA.
E-mail: ronald@biostat.wisc.edu

0035–9254/02/51485

Kaplan–Meier estimates with right-censored data (Kaplan and Meier, 1958). If the empirical CDFs (or Kaplan–Meier estimates) do not satisfy the ordering constraint, it may be desirable to use an alternative estimator that satisfies the constraints.

The most common approach to estimation subject to a stochastic ordering constraint is nonparametric maximum likelihood estimation. Brunk *et al.* (1966) derived the nonparametric maximum likelihood estimator (NPMLE) of two stochastically ordered cumulative distribution functions with complete or right-censored data. Dykstra (1982) demonstrated that the NPMLE has the same form as a product limit estimator based on modified data. He exploited this property to develop a simple algorithm for calculating the NPMLE and to demonstrate its consistency. Feltz and Dykstra (1985) and Dykstra and Feltz (1989) proposed algorithms for calculating the NPMLE of more than two distribution functions under stochastic ordering based on iteratively solving the pairwise problems. Hoff (2000) described an alternative algorithm for calculating the NPMLE based on the EM algorithm.

One objection to the NPMLE is that any violation of the stochastic ordering in the empirical CDFs, even at a single point, can and frequently does result in large differences between the NPMLEs and the empirical CDFs even in sections of the data where the empirical CDFs satisfy the constraints. Lo (1987) proposed a simple alternative to the NPMLE for the two-sample problem. Lo's estimator is obtained by swapping the survival function (or CDF) values between samples when the constraints are violated. Rojo and Ma (1996) conducted a simulation study comparing Lo's estimator with the NPMLE and found substantial improvements in both the bias and the mean-squared error.

In this paper, we propose to estimate the survival functions of k stochastically ordered random variables by using minimum distance estimation (Wolfowitz, 1957). The estimator proposed can be found by using a bivariate isotonic regression algorithm (Qian and Eddy, 1996). In Section 2, we describe both the proposed minimum distance estimator (MDE) as well as a straightforward extension of Lo's estimator to more than two samples. In Section 3, we present simulation results comparing the MDE with the NPMLE and the extension of Lo's estimator in terms of the bias and mean-squared error. In Section 4, we illustrate the application of the MDE to the assessment of reproducibility of macular edema grading.

2. Methodology

Consider the problem of estimating k cumulative distribution functions $F_1, F_2, \ldots, F_k$ subject to the stochastic ordering constraint that, for all x, $F_1(x) \geqslant F_2(x) \geqslant \ldots \geqslant F_k(x)$. (Here, population 1 is stochastically smaller than population 2, and so on.) Denote the empirical CDFs by $F_1^*, F_2^*, \ldots, F_k^*$. The empirical CDFs may not satisfy the restrictions desired. Intuitively, an estimator satisfying the stochastic ordering should be chosen to be close (in some sense) to the empirical CDFs. In the literature, estimates based on this criterion are called minimum distance estimates (Wolfowitz, 1957). These estimates have typically been used in parametric estimation problems (Parr, 1985). In this paper, we discuss how such estimates may be developed for a nonparametric estimation problem subject to constraints.

One measure of 'distance' between distribution functions is the Cramér–von Mises distance from F to G (Pettitt, 1982). Since one-to-one monotone transformations of the observed data preserve the underlying structure of our estimation problem (i.e. the stochastic ordering constraint), it would be desirable for our distance measure to be invariant under such transformations. If F is not continuous, the Cramér–von Mises distance is invariant only for monotone increasing transformations and not for monotone decreasing transformations.

To create a distance measure that is invariant under all monotone transformations, we denote the left continuous versions of the CDFs by F^- and G^-, i.e., if X is a random variable from distribution F, then $F^-(x) = \Pr(X < x)$. Define the distance from F to G by

$$d(F,G) = \frac{1}{2}\left[\int \{G(x) - F(x)\}^2 \,\mathrm{d}F(x) + \int \{G^-(x) - F^-(x)\}^2 \mathrm{d}F(x)\right].$$

This distance measure reduces to the Cramér–von Mises distance if F is continuous and is invariant under any one-to-one monotone transformation of the underlying data for any pair of distribution functions. Thus, it represents a suitable extension of the Crámer–von Mises distance for our purposes.

To estimate $F_1, F_2, \ldots, F_k$, we propose to use the distribution functions $G_1, G_2, \ldots, G_k$ satisfying the stochastic ordering constraint that minimize the following weighted sum of distances from the empirical CDFs to the estimated CDFs:

$$\sum_{i=1}^{k} w_i\, d(F_i^*, G_i).$$

We weight the distance for each sample by its corresponding sample size. With right-censored data, we replace the empirical CDFs by the Kaplan–Meier estimates and use the observed number of events in each population as weights.

If $x_1 < x_2 < \ldots < x_m$ are the observed data values and $n_{i,j} = w_j\{F_j^*(x_i) - F_j^*(x_{i-1})\}$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, k$, then the weighted criterion is

$$\sum_{j=1}^{k} w_j\, d(F_j^*, G_j) = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{n_{i,j} + n_{i,j+1}}{2}\{G_j(x_i) - F_j^*(x_i)\}^2.$$

The minimizing CDFs $G_1, G_2, \ldots, G_k$ subject to the stochastic ordering constraint can be found by using a bivariate isotonic regression algorithm such as the sandwich isotonic block class algorithm of Qian and Eddy (1996). Under very minimal conditions, the MDEs are strongly uniformly consistent. A proof of this result is achieved by noting that strong uniform consistency is equivalent to convergence of the modified Cramér–von Mises distance to 0, which is easily established (assuming that the weights are bounded away from 0 in the limit).

An alternative simple estimator can be obtained by extending the two-sample estimator of Lo (1987) to three or more samples as follows. For each x, take $H_j(x) = F_{(k-j)}^*(x)$ as the estimator of $F_j(x)$ for $j = 1, 2, \ldots, k$, where $F_{(1)}^*(x) \leqslant F_{(2)}^*(x) \leqslant \ldots \leqslant F_{(k)}^*(x)$ are the ordered empirical CDF values. It follows directly from the strong uniform and mean-squared error consistency of the empirical CDFs that this extension of Lo's estimator is both strongly uniformly consistent and mean-squared error consistent. The argument parallels the argument in Rojo and Ma (1996) for the two-sample case.

For characterizing the limiting behaviour of the estimators, we must distinguish between cases in which the ordering constraints are strict (i.e. $F_j(x) > F_{j+1}(x)$ for all x and$j = 1, 2, \ldots, k-1$) and cases in which the ordering constraints are not strict (i.e. $F_j(x) = F_{j+1}(x)$ for some x and$j$). If the ordering constraints are strict, the limiting distributions for the constrained estimators (the NPMLE, MDE and Lo's estimator) are identical with the limiting distributions for the empirical CDFs (or the Kaplan–Meier estimates). If the constraints are not strict, the limiting distributions of the constrained estimators are non-normal. See Præstgaard and Huang (1996) for a discussion of the limiting distribution of the NPMLE in the two-sample setting. We would follow Præstgaard and Huang (1996) in applying the standard variance formulae for the empirical CDFs ($F(1 - F)/n$) or Kaplan–Meier estimators (Greenwood's formula) to obtain conservative pointwise confidence bounds.

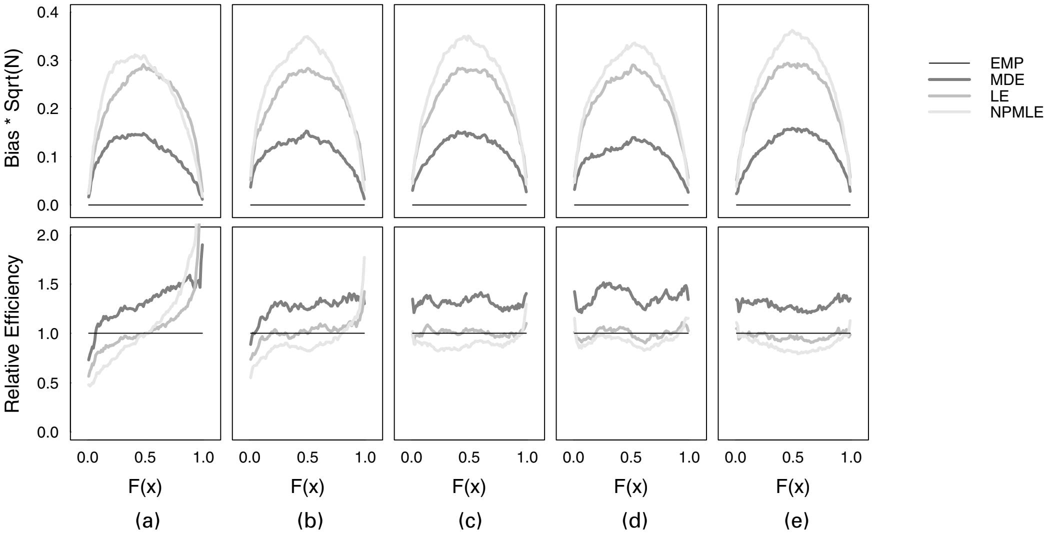
3. Simulation results

We conducted a simulation study to evaluate the bias and mean-squared error functions for the NPMLE, Lo's estimator and its extension LE, and the proposed MDE as well as for the empirical CDFs EMP. A total of 1000 simulations were performed for each experiment. The code was written in S-PLUS and Fortran. Hoff's algorithm was used to calculate the NPMLE estimator.

As noted by Præstgaard and Huang (1996), the most important case for understanding the behaviour of order-restricted estimators is the case in which all constraints are active, i.e. $F_1 = F_2 = \ldots = F_k \equiv F$. In this case, we can, without loss of generality, restrict attention to the uniform distribution.

For this simulation study, we restrict our attention to the two-sample problem. Fig. 1 presents the bias and relative efficiency (mean-squared error relative to the true mean-squared error of the empirical CDF) of the four estimators for sample sizes per group of 10, 100, 1000, 10000 and 100000. Results are presented for the population assumed to be stochastically smaller; given the symmetry of the problem, results from the other population must be similar. In Fig. 1, we observe that the bias of the NPMLE is substantially larger than the bias of the extension of Lo's estimator, which, in turn, is substantially larger than the bias of the MDE regardless of the sample size. This ordering of the estimators holds for all sample sizes, although the magnitude of the bias is reduced for larger sample sizes.

The MDE produces an increase in efficiency of approximately 30% over the empirical CDFs. The extension of Lo's estimator and the NPMLE generally do not produce any significant gain in efficiency over the empirical CDFs. In general, the NPMLE is the least efficient estimator. On the basis of these simulation results, our preferred estimator for enforcing the ordering constraint would be the MDE, which has the smallest bias and demonstrates substantial improvements in efficiency. Moreover, there appears to be little reason to prefer either the extension of Lo's estimator or the NPMLE over the empirical CDFs.



(a) (b) (c) (d) (e)

Fig. 1. Bias and relative efficiency (mean-squared error relative to that of the empirical CDF) functions of the NPMLE, the extension of Lo's estimator LE, the MDE and the empirical CDFs EMP based on samples of size (a) 10, (b) 100, (c) 1000 (d) 10000 and (e) 100000 from two identical uniform distributions (results are presented for the population assumed to be stochastically smaller and are based on 1000 simulations)

4. Example: nuclear sclerosis grading

The case-study originates from the development of a grading scale for nuclear sclerosis, which has been a focus of study at the Fundus Photograph Reading Center at the University of Wisconsin—Madison. Nuclear sclerosis, a form of cataract, is currently assessed by means of a set of six standard photographs. The level of nuclear sclerosis is determined in a 'paint strip' method. A photograph is first placed within the set of six standard photographs, and then an appropriate decimalized score is assigned. For example, a photograph falling between standards 2 and 3 and judged to be slightly closer to standard 2 might be assigned a value of 2.3.

Our goal is to characterize the underlying variability of the nuclear sclerosis gradings. For this purpose, the Reading Center has accumulated a collection of 966 photographs, each of which has been assessed twice (i.e. each photograph has been sent through the grading system twice). As a characterization of the variability in nuclear sclerosis gradings, we choose the distribution of a replicate grade (y) conditional on the original grade (x), which is easily interpreted. For example, a question of great interest to graders and researchers alike is, given an original nuclear sclerosis grade of 2.0, what is a plausible range of grades if the photograph were re-graded? Our approach answers this question by producing different (and likely asymmetric) bounds for each original grade. As an alternative, the Bland–Altman approach (Bland and Altman, 1986) provides a global bound on the $|y-x|$ that is applicable to all levels of severity and is most useful if the resulting bounds indicate a negligible difference. In our setting, where grading variability is non-negligible and non-constant, the full conditional distribution is of greater use.

Since nuclear sclerosis gradings are subjective assessments along a potentially unequally spaced scale, it is desirable to require as few assumptions regarding the conditional distributions as possible. In many settings, particularly for scales measuring the severity of lesions known to be present, a simple and natural restriction for grading data is to assume stochastic ordering, i.e. the distribution of a replicate grade given a higher original grade should be stochastically larger than the distribution of a replicate grade given a lower original grade.

We estimate the $k=50$ conditional distributions for replicate grades given an original grade by using the MDE described earlier. A calculation of the NPMLE is not feasible for 50 populations by using Hoff's (2000) algorithm. Given the poor performance of the NPMLE in other settings, we chose not to pursue an alternative algorithm for its calculation.

To see this, consider photographs that are originally assigned grades of x_1 and x_2 ($x_1 < x_2$). If the grading system is actually measuring the severity of disease, the true level of severity for a photograph assigned grade x_1 should, on average, be lower than the true level of severity for the photographs assigned grade x_2. Thus, repeat grades for photographs initially assigned grade x_1 should not, on average, be lower than repeat grades for photographs initially assigned grade x_2. Similarly, if we dichotomize the repeat grades (above or below a threshold value), we would expect a higher proportion of photographs initially graded as x_1 to fall above the threshold than of photographs initially graded as x_2. But, as noted in Section 1, this is simply a restatement of the stochastic ordering assumption. This assumption may be unreasonable in settings where the primary differences in gradings relate to the existence of a lesion rather than to the severity of the lesion. This concern does not apply to grading nuclear sclerosis as its presence is certain and only its severity is in question.

In Fig. 2, we display the variability of the nuclear sclerosis gradings in two fashions. In the available data set, the designation of original and replicate gradings is arbitrary, so we utilized both possible designations of each pair of gradings in our analysis. Fig. 2(a) displays all pairs of gradings (jittered so duplicate points are visible) along with selected percentiles of the estimated

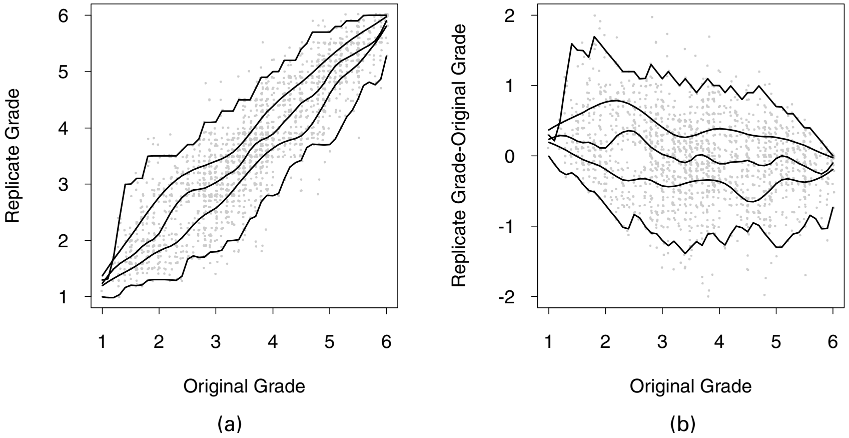


Fig. 2. Variability of replicate nuclear sclerosis gradings: (a) replicate grade *versus* original grade; (b) difference between grades *versus* original grade (•, observed pairs of gradings; ———, smoothed versions of 2.5th, 25th, 50th, 75th and 97.5th percentiles from the minimum distance estimates of the distribution functions)

distributions (MDEs subject to stochastic ordering) of a replicate grade for each original grade. Fig. 2(b) displays the same information but uses the difference between gradings for the y-axis instead of the replicate grading. For display, the percentiles are smoothed by using the S-PLUS function `smooth.spline` with the smoothing parameter chosen by using generalized cross-validation. Owing to the rarity of extremely low or high grades, estimates at either extreme are not reliable and should be discounted.

Fig. 2 presents many interesting features of the variability of the underlying data. First, it provides a general picture of the variability that is inherent in the practice of assessing photographs. Even researchers with little statistical background can examine the 95% prediction bounds and acquire a basic understanding of the underlying variability in the grading system. Providing this knowledge to researchers who are establishing safety guidelines and progression thresholds will help them to establish a range where a signal can be safely seen through the noise generated from the difficulty of assessing these photographs.

For example, a closer look at Fig. 2 shows that variability is neither constant nor symmetric across the scale; grades around standard 2 have a higher level of variability than those around standard 5. For an original grade of 2.0, a 95% equal-tailed prediction interval is 1.3–3.5; for an original grade of 5.0, a 95% equal-tailed prediction interval is 4.5–5.5. This observation may indicate that the graders find it more difficult to distinguish photographs near standards 2 and 3 than photographs near standards 4 and 5. There are at least two possible explanations for this. First, the graders might not have been adequately trained to use that portion of the scale or might have forgotten that training. If this is so, the Reading Center could conduct additional training focusing on those portions of the scale that are yielding more discordant assessments. It is worth noting, however, that the apparent problem area is the most frequently used portion of the scale, making the hypothesis that the graders forgot how to grade that portion of the scale less tenable.

However, these findings may simply reflect the fact that the standard photographs are (and hence the grading scale is) not ‘equally spaced’. One proposal that is currently under consideration by the Reading Center (which is not based on our analysis) is to rescale the grades so that the distance between standards 1 and 3 is equal to 1 unit. Under this new scale, the grades for

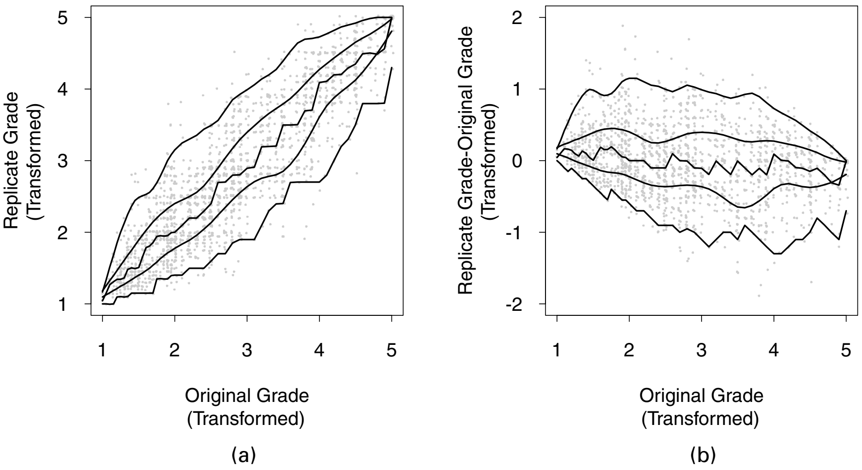


Fig. 3. Variability of replicate nuclear sclerosis gradings using a transformed scale (standard 1 ≡ 1; standard 2 ≡ 1.5; standard 3 ≡ 2; standard 4 ≡ 3; standard 5 ≡ 4; standard 6 ≡ 5): (a) replicate grade *versus* original grade; (b) difference between grades *versus* original grade (•, observed pairs of gradings; ———, smoothed versions of the 2.5th, 25th, 50th, 75th and 97.5th percentiles from the minimum distance estimates of the distribution functions)

the six standard photographs would be 1, 1.5, 2, 3, 4 and 5. Since the transformation between scales is monotone, the estimates for the proposed new scale can be found by simply applying the transformation to our estimates for the original scale. In Fig. 3, we display the variability of replicate gradings by using this modified scale. As we might have expected, the variability under this transformation is relatively constant across the entire scale. For an original grade of 1.5 on the transformed scale (2.0 on the untransformed scale), the 95% equal-tailed prediction interval is 1.15–2.50; for an original grade of 4.0 on the transformed scale (5.0 on the original scale), the 95% equal-tailed prediction interval is 3.50–4.50. For a wide range of original grades (excluding the extremes), the upper limit on the 95% prediction interval on the transformed scale is roughly 1.0 unit above the original grade.

5. Discussion

Minimum distance estimation provides a reasonable alternative to nonparametric maximum likelihood estimation for stochastically ordered distributions. Although both estimators are consistent, for a wide range of sample sizes, the MDE is superior, in terms of both the bias and the mean-square error. In practice, the MDE preserves the unrestricted NPMLE unless the constraints are violated, whereas the restricted NPMLE can drastically alter the estimates, even in regions where the unrestricted NPMLE satisfies the ordering constraint. We demonstrated the usefulness of the methodology in assessing the variability of repeat grading, particularly in settings such as nuclear sclerosis grading where the grading variability is non-constant and substantial.

Acknowledgements

The authors would like to thank Murray Clayton and Tom Cook of the University of Wisconsin for their helpful comments on an earlier draft of this paper. The authors also wish to thank the Joint Editor, Associate Editor and two referees for several suggestions which have greatly improved the paper. The research was partially supported by National Institutes of Health grant EY 07119.

References

Bland, J. M. and Altman, D. G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, i, 307–310.

Brunk, H. D., Franck, W. E., Hanson, D. L. and Hogg, R. V. (1966) Maximum likelihood estimation of the distributions of two stochastically ordered random variables. *J. Am. Statist. Ass.*, **61**, 1067–1080.

Dykstra, R. L. (1982) Maximum likelihood estimation of the survival functions of stochastically ordered random variables. *J. Am. Statist. Ass.*, **77**, 621–628.

Dykstra, R. L. and Feltz, C. J. (1989) Nonparametric maximum likelihood estimation of survival functions with a general stochastic ordering and its dual. *Biometrika*, **76**, 331–341.

Feltz, C. J. and Dykstra, R. L. (1985) Maximum likelihood estimation of the survival functions of *n* stochastically ordered random variables. *J. Am. Statist. Ass.*, **80**, 1012–1019.

Hoff, P. D. (2000) Constrained nonparametric maximum likelihood via mixtures. *J. Comput. Graph. Statist.*, **9**, 633–641.

Kaplan, E. L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Statist. Ass.*, **53**, 457–481.

Lo, S. (1987) Estimation of distribution functions under order restrictions. *Statist. Decsns*, **5**, 251–262.

Parr, W. C. (1985) Minimum distance estimation. In *Encyclopedia of Statistical Science*, vol. 5, pp. 529–532. New York: Wiley.

Pettitt, A. N. (1982) Cramér–von Mises statistic. In *Encyclopedia of Statistical Science*, vol. 2, pp. 220–221. New York: Wiley.

Præstgaard, J. T. and Huang, J. (1996) Asymptotic theory for nonparametric estimation of survival curves under order restrictions. *Ann. Statist.*, **24**, 1679–1716.

Qian, S. and Eddy, W. F. (1996) An algorithm for isotonic regression on ordered rectangular grids. *J. Comput. Graph. Statist.*, **5**, 225–235.

Rojo, J. and Ma, Z. (1996) On the estimation of stochastically ordered survival functions. *J. Statist. Computn Simuln*, **55**, 1–21.

Wolfowitz, J. (1957) The minimum distance method. *Ann. Math. Statist.*, **28**, 75–88.