# Small Sample Comparison of Five Bivariate Survival Curve Estimators

By

Ronald C. Pruitt

Technical Report No. 559

School of Statistics

University of Minnesota

4 March 1991

[1]Research partially supported by National Institutes of General Medical Sciences Grant GM-42921-01.

## Abstract

In this paper we study the performance of five proposed bivariate survival curve estimators. The estimators are those of Muñoz (1980), Langberg and Shaked (1982) and Campbell and Földes (1982), Tsai, Leurgans, and Crowley (1986), Dabrowska (1988), and Pruitt (1991a). The performance of the estimators is compared for data generated by three bivariate models: the bivariate exponential model of Marshall and Olkin (1967), the bivariate exponential model of Downton (1970), and a linear combination of exponentials model. The estimators are compared in their ability to estimate correlations, survival functions, and probabilities.

## 1 Introduction.

We consider the problem of estimating a bivariate survival curve with bivariate right censored data. Many estimators have been proposed for this problem, including those of Muñoz (1980), Campbell (1981), Langberg and Shaked (1982), Campbell and Földes (1982), Hanley and Parnes (1983), Tsai, Leurgans, and Crowley (1986), Dabrowska (1988), and Pruitt (1991a, 1991b).

Muñoz (1980), Campbell (1981), and Hanley and Parnes (1983) all discuss the generalized maximum likelihood estimator (GMLE) for this problem. This estimator is known to be consistent if the distribution being estimated is purely discrete (Campbell, 1981), but may be inconsistent for continuous data (see Leurgans, Tsai, and Crowley, 1982). The estimator is also not unique for samples taken from absolutely continuous distributions. Recently, a smoothed version of this estimator has been proposed (Pruitt, 1991b).

Langberg and Shaked (1982) and Campbell and Földes (1982) decompose $\Pr[X_1 > x_1, X_2 > x_2]$ as $\Pr[X_2 > x_2 | X_1 > x_1] \Pr[X_1 > x_1]$ and estimate each term separately. The resulting estimates are usually not proper survival functions and depend on the ordering of the decomposition.

Tsai, Leurgans, and Crowley (1986) propose an estimator, here called the TLC estimator, using nonparametric smoothing techniques relying on a decomposition of a bivariate survival function, and show it to be uniformly consistent. Although nonparametric smoothing techniques are used, it seems likely that the estimator of the survival function converge at the usual $n^{-1/2}$ rate, just as an integrated density estimator converges faster asymptotically than the density estimator itself. Some negative features of the TLC estimator are that it is not affine equivariant and only assigns mass to uncensored observations and singly censored observations with the uncensored value smaller than the censored value.

Dabrowska (1988) estimates components of the bivariate cumulative hazard function separately and uses a product limit form for the survival curve as a function of these quantities. Her estimator is also usually not a proper survival function.

Pruitt (1991a) uses nonparametric smoothing techniques in a different way to get a bivariate survival curve estimator. He uses a localized product limit estimator to impute values for singly censored observations and then uses generalized maximum likelihood ideas to complete the estimator.

## 2 Background.

Let the unobservable survival times of interest be given by $(X_1, X_2)$ and a nuisance censoring variable be given by $(C_1, C_2)$. We consider bivariate right censoring, where the observable variables are $Y_1, Y_2, D_1$, and $D_2$, where $Y_j = X_j \wedge C_j$ and $D_j = 1[Y_j = X_j]$. It has usually been assumed that $(X_1, X_2)$ and $(C_1, C_2)$ are independent to ensure the identifiability of the survival function. This assumption is stronger than necessary for identifiability (see Pruitt, 1990), but all the parametric models considered follow it. The goal is to estimate the survival function for $(X_1, X_2)$, $\bar{F}(x_1, x_2) = P[X_1 > x_1, X_2 > x_2]$.

## 3 Estimators.

We do not give details of the various estimators but rather refer readers to the original papers cited. Here we describe some of the specific implementation details used in this study. All of the estimators described here can have some indeterminacy at the largest values of the data set under certain conditions, similar to that in the univariate case when the largest observed value is censored. In these cases mass was placed as close to the origin as possible. All of the computations were done with C programs that are capable of being run under S. These routines are publicly available, for details, see Appendix A.1.

**Pruitt.** The smoothing kernel used was a uniform kernel. The half window widths used were 1.2, 1.0, and 0.8 for sample sizes 10,25, and 50. From now on we will refer to the half window width as the window width. No attempt was made to optimize either the form of the kernel or the window width.

**Dabrowska.** There were no special considerations for the Dabrowska estimator other than those mentioned in the general introduction.

**GMLE.** The GMLE is generally not unique. A GMLE was computed by iterating the EM algorithm from a specific starting point, which is described in Appendix A.2.

**TLC.** The smoothing kernel used was a uniform kernel. The widths were the same as used for the Pruitt estimator.

**Pathwise.** The pathwise estimator (Langberg and Shaked, 1982; Campbell and Földes, 1982) was computed as the average of the estimators obtained from the two decompositions, $P[X_1 > x_1, X_2 > x_2] = P[X_1 > x_1 | X_2 > x_2] P[X_2 > x_2]$ and $P[X_1 > x_1, X_2 > x_2] = P[X_2 > x_2 | X_1 > x_1] P[X_1 > x_1]$. The software is capable of giving any convex combination of the two pathwise estimators.

# 4 Models.

We consider data generated according to three different models which are briefly described below.

**Marshall-Olkin.** The bivariate exponential model of Marshall and Olkin is given by

$$\bar{F}(x_1, x_2) = P[X_1 > x_1, X_2 > x_2] = \exp\{-\lambda(x_1 + x_2) - \Lambda(x_1 \vee x_2)\},$$

for $0 < x_1, x_2, \lambda, \Lambda$. We only consider common scale for $X_1$ and $X_2$ as the estimators are scale equivariant with the exception of the TLC estimator. The model may be described by supposing that failures are caused by three types of shocks on a system with two components. These mutually independent shocks occur at rates $\lambda, \lambda$, and $\Lambda$, and cause failures of the first, second, or both components. If the shocks follow Poisson distributions, the model above results. The correlation between $X_1$ and $X_2$ may be shown to be $\rho = \Lambda/(2\lambda + \Lambda)$. This model gives nonnegative correlations, but the correlation structure is very specialized. Conditional on $X_1 \neq X_2$, $X_1$ and $X_2$ are independent.

The integrated squared error,

$$\int\int (\bar{F}(s,t) - \hat{\bar{F}}(s,t))^2 ds dt, \tag{4.1}$$

may be obtained for step function estimators $\hat{\bar{F}}$, which is an extra measure of comparison of the estimators which is not available for the other two models. This calculation is described in Appendix A.3. The values of the survival and distribution functions are also available analytically.

**Downton.** The model of Downton is given by the following. Suppose $R$ is geometric with parameter $1-\rho$:

$$f(r) = \rho^{r-1}(1-\rho) \qquad\qquad r = 1, \ldots,$$

and conditional on $R = r$, $X_1$ and $X_2$ are independently gamma distributed with parameters $r$ and $\lambda/(1-\rho)$:

$$f(x) = \frac{\lambda^r x^{r-1}}{\Gamma(r)(1-\rho)^r} \exp\{-\lambda x/(1-\rho)\}.$$

The marginal distribution of $X_i$ may be checked to be exponential with parameter $\lambda$ and the correlation between $X_1$ and $X_2$ is $\rho$. There is no closed form for the survival or distribution functions and so values of these used in the simulations were computed by Monte Carlo. The correlation structure is not as rigid as that imposed by the Marshall and Olkin model.

**Linear combination of exponentials.** This model is given by the following. Let $T_1, T_2$, and $T_3$ be i.i.d. exponential with parameter lambda. Let $0 \le \alpha \le 1$, and let $\beta = (\alpha^2 + (1-\alpha)^2)^{-1/2}$. Then let

$$X_i = \alpha\beta T_3 + (1-\alpha)\beta T_i + (1-\beta)\lambda^{-1}.$$

This gives the marginal distributions the same mean and variance as exponential with parameter $\lambda$. The correlation between $X_1$ and $X_2$ is $\rho = \alpha^2/(\alpha^2 + (1-\alpha)^2)$. The model as stated can give rise to observations with negative values, and since the software is set up to handle nonnegative values only, the values were shifted to be positive.

## 5 Results.

The three models we are considering are all parametrized by $\lambda$ and $\rho$. We will denote the parameters for the $(X_1, X_2)$ distribution by $(\lambda_x, \rho_x)$ and for the $(C_1, C_2)$ distribution by $(\lambda_c, \rho_c)$. In the simulations we always take $\lambda_x = 1$. The simulation study consisted of generating 10000 samples (2500 for $n = 50$) under each combination of the following conditions:

- 3 models: Marshall-Olkin, Downton, linear combination
- 4 $\rho_x$ values: 0, 0.2, 0.333, 0.5

- 3 $\lambda_c$ values: 0.5, 1, 2
- 4 $\rho_c$ values: 0, 0.2, 0.333, 0.5
- 3 sample sizes: 10, 25, 50 .

For each sample, the five estimators were computed and the survival probability $\bar{F}(\log 2, \log 2) = P[X_1 > \log 2, X_2 > \log 2]$, the probability $F(\log 1.5, \log 1.5) = P[X_1 \leq \log 1.5, X_2 \leq \log 1.5]$, and the correlation were estimated(by the correlation of the estimated distribution). The true values are known for the Marshall-Olkin model, and the survival and distribution function probabilities were estimated by 1,000,000 Monte Carlo trials for the other two models. For the Marshall-Olkin model the integrated squared error was also computed. The biases and mean squared errors were estimated for the survival,probability, and correlation problems. The number of points assigned mass by the estimators was also noted. An example of the available output for all variables independent unit exponential ($\rho_x = 0, \lambda_c = 1, \rho_c = 0$, and $n = 50$) is given in Table 1.

### 5.1 Comparison of all five estimators.

The conditions for the Marshall-Olkin model are somewhat different than for the other two models. The other two models have mean $\lambda$ and correlation $\rho$ if these are the parameters of the model, but this is not true for the Marshall-Olkin model. For the Marshall-Olkin model the means are 1, 0.667, 0.5, and 0.333 for correlations 0, 0.2, 0.333, and 0.5. The correlated variables tend to have smaller means for the same $\lambda$ parameter. This makes the conditions with $\lambda_x = 1 \rho_x = 0, \lambda_c = 2, \rho_c = 0.5$ very difficult indeed. It may be computed that the probability of getting an uncensored observation is 1/15, a singly censored observation 8/70 (for each type), and a doubly censored observation 74/105. We only expect about 3 uncensored observations in a sample of size 50.

After the main study was completed, and it was noted that the weakness of the Pruitt estimator was under extreme conditions of censoring, and even more extreme test was conducted. The distribution for $\vec{X}$ was an equal mixture of three bivariate normals: the first ($\mu_1 = .45, \mu_2 = .25, \sigma_1 = .149, \sigma_2 = .112, \rho = -0.722$), the second ($\mu_1 = 1, \mu_2 = .8, \sigma_1 = .3, \sigma_2 = .101, \rho = -0.197$), and the third ($\mu_1 = 1, \mu_2 = 1.2, \sigma_1 = .112, \sigma_2 = .304, \rho = 0.588$).

| Problem | | IMSE | | Correlation | | |
|---|---|---|---|---|---|---|
| Estimator | Average size | Average | Relative efficiency | Bias | MSE | Relative efficiency |
| Pruitt | 147.6 | 0.0343(0.0006) | 1.00 | -0.017 | 0.0655(0.0028) | 1.00 |
| Dabrowska | 569.1 | 0.0433(0.0010) | 0.79 | 0.004 | 0.0774(0.0036) | 0.85 |
| GMLE | 77.8 | 0.0383(0.0008) | 0.90 | -0.003 | 0.0488(0.0020) | 1.34 |
| TLC | 63.2 | 0.0559(0.0012) | 0.61 | -0.061 | 0.0760(0.0030) | 0.86 |
| Pathwise | 1106.8 | 0.0472(0.0008) | 0.73 | -0.179 | 0.1151(0.0042) | 0.57 |

| Problem | Survival | | | Probability | | |
|---|---|---|---|---|---|---|
| Estimator | Bias | MSE | Relative efficiency | Bias | MSE | Relative efficiency |
| Pruitt | -0.003 | 0.0054(0.0002) | 1.00 | 0.001 | 0.0023(0.0001) | 1.00 |
| Dabrowska | 0.003 | 0.0069(0.0003) | 0.78 | 0.001 | 0.0029(0.0001) | 0.80 |
| GMLE | -0.024 | 0.0062(0.0003) | 0.87 | 0.011 | 0.0036(0.0002) | 0.64 |
| TLC | 0.003 | 0.0097(0.0005) | 0.55 | 0.001 | 0.0029(0.0001) | 0.79 |
| Pathwise | 0.002 | 0.0081(0.0004) | 0.66 | 0.001 | 0.0037(0.0002) | 0.64 |

Table 1: Example output with all times independent exponential. Standard errors are in parentheses. Efficiencies are relative to the Pruitt estimator.

The second and third distributions overlap but are fairly distinct from the first. $C_1$ and $C_2$ were independent unit exponential. The expectation was that smoothing, and hence the Pruitt estimator, might perform poorly in this environment since the $X$ distribution is unsmooth compared with the $C$ distribution. The survival probability $\bar{F}(1,1)$, the probability $F(1.1, 0.9)$, and the correlation were computed for each estimator. The true values were approximated from 1,000,000 Monte Carlo trials (although the correlation may be obtained analytically). The results for this trial are given in Table 13 and are discussed more fully below.

A feature which is not stressed in the following analysis is the bias of the various estimators. The negative effects of bias are already incorporated into the mean squared error. It may be noted by examining Tables 10 - 13 or the original output that Dabrowska's estimator tends to be the least biased, although all the estimators have bias small enough so as to contribute little to the mean squared error.

**IMSE problem.** Here the results were amazingly constant over all the different conditions examined. The Marshall-Olkin model was the only one for which the IMSE could be computed. The IMSE reflects the ability to estimate the survival function over the entire range of survival times rather than concentrating on a single point as in the survival problem. But from the survival problem we may be able to gain some insight into how applicable the results are to the other types of models discussed here. In particular by examining the survival problem results for the different models in Figures 3 and 4 we see that the GML estimator does better at the MO model than at the other two while the results are similar across models for the other four estimators. We expect that the IMSE results may make the GML appear slightly better than if we had been able to conduct IMSE results over all three models.

The most surprising thing about the IMSE results is their constancy over all the 48 conditions examined. The rankings almost always occured in the order: Pruitt, GML, Dabrowska, pathwise, and TLC. Relative efficiencies also showed no extremely marked trends marginally over each of the conditions examined. There are however interactions between the conditions examined, see the paragraph on interactions. The largest trend is for varying the correlation between the censoring variables; all the estimators improve in relative efficiency compared to the Pruitt estimator as this correlation increases. There is a smaller effect of decreasing relative efficiency for the four non-Pruitt estimators as the correlation of the survival times is increased.

The only conditions under which the relative rankings of the various estimators change are the following. For 3 combinations of conditions the TLC and pathwise estimator switch so that the TLC is better if the $X$ correlation is low, the $C$ correlation is high, and $\lambda_c$ is low, that is, there is not much censoring. Under two conditions the ranking at the top is GML, Dabrowska, Pruitt; if the $X$ correlation is low, the $C$ correlation is high, and $\lambda_c$ is high so there is a lot of censoring. Under 4 other similar conditions the GML and Pruitt estimators change rankings so that the GML estimator has a lower IMSE. The Pruitt estimator seems to perform relatively worse under the very extreme conditions examined in which there was a lot of censoring.

Overall the relative efficiencies and especially the rankings stayed amazingly constant over the conditions examined. In terms of IMSE and the Marshall-Olkin model it is fairly easy to rank these estimators: Pruitt, GML, Dabrowska, pathwise, and TLC.

**Correlation problem.** The correlation problem turned out to be quite different than the other problems. From examining Figures 3 and 4 we can see that the relative ranking of the estimators is different and the efficiencies vary much more widely. In estimating the correlation, the GML is most efficient under the conditions examined followed by Pruitt, then TLC and Dabrowska closely grouped, and finally pathwise. The GML is the clear winner here with efficiency gains ranging from modest to over 50% over the next closest competitor depending on the conditions. From Table 7, the GML ranked first in 83% of the conditions examined. The Pruitt estimator and perhaps the TLC estimator would have done better with a better choice of kernel. The Pruitt estimator was most sensitive to kernel choice in the correlation problem, see Subsection 5.2.

The clear loser was the pathwise estimator which was very biased and took on values outside the interval $[-1, 1]$. As an extreme example, the bias of the pathwise estimator for $\rho_x = 0, \rho_c = .5, \lambda_c = 2$, and the Marshall-Olkin model was -0.85 with a mean squared error of 2.08 when the GML had a MSE of 0.129. These extreme readings are indicative of the problems with the pathwise estimator for this problem. The fact that the pathwise estimator is not a proper survival function seems to cause serious difficulty in estimating the correlation.

The TLC estimator may deserve more exploration for estimating correlations. Here it did not perform well, but if the kernel could be optimized it might perform better. The TLC estimator was the only one worth anything in the very challenging mixture of normals problem for estimating the correlation, see Table 13.

**Survival problem.** The survival problem shares many features with the IMSE problem. The average efficiencies depicted in Figure 4 are not much different than for the IMSE problem although those for pathwise and TLC are somewhat lower. We do see that the GML is better for the Marshall-Olkin model than for the other two models. These averages are hiding more variability than was the case for the IMSE problem. In particular the relative rankings are not as constant and GML and Dabrowska change places much more often. This is partly due to the different nature of the problem and partly due to examining more models. For the Marshall-Olkin model, Dabrowska, TLC, and GML have better average ranks for the survival problem than for the IMSE problem under the same conditions. This may be due to better estimation at the central value examined than closer to the edges of the space or for other reasons. The relative behavior at centrally placed values and at values nearer the extremes of survival was not directly examined in this study. But behavior for the survival problem is not exactly the same as for the IMSE problem even under similar conditions. A bigger effect of differences in the survival problem is the different models examined. The GML does better for the Marshall-Olkin model than for the other two models, and the Dabrowska estimator does better than the GML for the other two models examined in terms of relative rank, although they both trail Pruitt. The GML also does better at high values of $\lambda_c$ than Dabrowska. The overall ranking is the same as for the IMSE problem, although Dabrowska's showing is stronger in estimating one centrally located survival probability.

GML, Dabrowska, and Pruitt all performed similarly for the difficult mixture of normals problem. Pathwise was next followed distantly by TLC. Even under these conditions where it was expected that smoothing would prove difficult, the Pruitt estimator performed well.

**Probability problem.** The probability problem is not as clear cut as the other in terms of ranking all the estimators. The Pruitt estimator was clearly the best but the rankings of the others changed for different conditions. The GML did not perform as well as for the other problems, this may be due to the need to smooth the GML to get good performance in this problem, see Subsection 5.2. Overall, Dabrowska was second best although TLC and GML were competitive under certain conditions.

Pruitt, Dabrowska, and GML all performed similarly on the difficult mixture of normals problem. Pathwise and TLC did not perform well.

**Contour plots.** Contour plots were constructed for each of the estimators for a particular sample of size 50 obtained by taking $X_1, X_2, C_1$, and $C_2$ independent unit exponential. The results are given in Figure 6 along with a contour plot of the empirical survival function of the actual data set. It may be noticed that Dabrowska and pathwise are not survival functions. This is usually the case, and if this is a problem other estimators should be used. As we would expect the estimators show more agreement for the higher contours and less at low survival levels. In fact the Pruitt and Dabrowska estimators agree very well on the .6-.9 contours with the proviso that the Pruitt estimator is monotone. Whether this behavior occurs in general is unknown. Construction of such contour plots is easily accomplished using the S functions provided.

**Interactions.** As one assessment of the interactions between conditions on the various parameters, we performed an ANOVA decomposition of the relative efficiencies of the Dabrowska estimator compared to the Pruitt estimator for the IMSE problem. The results of the study, especially for the IMSE problem, were so clear that in general such an analysis would only serve to obscure the main points, and we have only included this one example to point out that the effects are not acting singly in these problems. The overall mean relative efficiency was 86.7%. The efficiency decreased in $\rho_x$, and increased in $\rho_c$ and $\lambda_c$ as may be seen in Figure 4. The coefficients are given in Table 2. The interactions between $\rho_x$ and $\rho_c$ were not significant. Increases in efficiency were associated with both $\rho_c$ and $\lambda_c$ being either small or large. The efficiency was smaller if one of $\lambda_c$ and $\rho_c$ was small while the other was large. The relationship between $\rho_x$ and $\lambda_c$ was the reverse: Efficiency was higher if only one was small or large, and efficiency was lower if both were small or large. The magnitude of these effects was similar to that of the main effects. All these effects on efficiency are positive and as large as possible for $\rho_x = 0, \rho_c = .5$, and $\lambda_c = 2$, and these conditions just managed to have the relative efficiency compared to the Pruitt estimator break 1. While the interactions are reinforcing under conditions which favor Dabrowska, the interactions are damping when Dabrowska is not favored. For example, if $\rho_c = 0$ and $\lambda_c = .5$ which are marginal conditions favoring Pruitt over Dabrowska, the interaction effect tends to dampen this: the main effects are -4.5 and -2.7 while the interaction effect is +4.4.

## 5.2 Choice of smoothing constants for the Pruitt and GML estimators.

General results on the choice of smoothing parameters for the Marshall-Olkin model are given in Figures 8 and 9. These figures show that averaged over all conditions the choice of smoothing constant is not particularly important for the Pruitt estimator, although small kernels can be somewhat inefficient. In particular the kernel for the correlation problem needs to be chosen sufficiently large or an efficiency penalty will be paid. The kernel actually used for the study was relatively efficient for the IMSE, survival, and probability problems but was only about 80% as efficient as the best kernel for the correlation problem. A larger kernel would have performed better for all of the problems. Figure 7 shows the optimal kernels for the Pruitt estimator under various conditions. It can be seen that the kernel used was generally smaller than the optimal kernel, often by a large amount. The optimal amount of smoothing is seen to decrease with increasing correlation of either the $X$ or $C$ variables as well as with increasing censoring. A more adaptive choice of window width may be preferable. But with the exception of the correlation problem, this misspecification of the kernel, which may be difficult to avoid in practice, did not cause much efficiency loss. The Pruitt estimator with a kernel not chosen particularly well was still able to perform more than adequately in this simulation study.

| $\rho_x = 0$ | $\rho_x = .2$ | $\rho_x = .333$ | $\rho_x = .5$ |
|---|---|---|---|
| 2.6 | -0.4 | -1.2 | -1.0 |

| $\rho_c = 0$ | $\rho_c = .2$ | $\rho_c = .333$ | $\rho_c = .5$ |
|---|---|---|---|
| -4.5 | -1.4 | 0.6 | 5.3 |

| $\lambda_c = .5$ | $\lambda_c = 1$ | $\lambda_c = 2$ |
|---|---|---|
| -2.7 | -2.5 | 5.2 |

| | $\rho_x = 0$ | $\rho_x = .2$ | $\rho_x = .333$ | $\rho_x = .5$ |
|---|---|---|---|---|
| $\lambda_c = .5$ | -3.6 | -0.7 | 0.5 | 3.8 |
| $\lambda_c = 1$ | 1.0 | -0.8 | -0.4 | 0.2 |
| $\lambda_c = 2$ | 2.6 | 1.6 | -0.2 | -4.0 |

| | $\rho_c = 0$ | $\rho_c = .2$ | $\rho_c = .333$ | $\rho_c = .5$ |
|---|---|---|---|---|
| $\lambda_c = .5$ | 4.4 | 1.1 | -1.6 | -3.9 |
| $\lambda_c = 1$ | -1.4 | -0.7 | 0.4 | 1.7 |
| $\lambda_c = 2$ | -3.0 | -0.4 | 1.2 | 2.2 |

Table 2: Anova decomposition coefficients for relative efficiency of Dabrowska estimator to Pruitt estimator in the IMSE problem.

For the much more difficult mixture of normals problem, the same qualitative behavior was observed. The relative efficiency compared to the optimal kernel was over .97 in the survival problem for kernels with width greater than 0.21, and was over .96 in the probability problem for these same kernels. For the correlation problem the efficiency dropped off somewhat more: kernel widths between 0.21 and 0.81 gave relative efficiencies over 0.9 and widths larger than 0.11 gave relative efficiencies over 0.8. As commented on before, all the estimators were essentially useless for estimating the correlation in this problem with the possible exception of TLC. It is dangerous to generalize from these results to the choice of smoothing constants for the Pruitt estimator in general, but in the problems examined the choice of smoothing constant was not particularly critical.

The results given for the GML estimator in Figure 9 show a different story for this estimator. The amount of smoothing examined was less, and more smoothing than the results shown was found to cause more efficiency loss. Here it is always best not to smooth at all, except in the probability problem. To get consistency of the estimator it is necessary to provide some smoothing for large enough sample sizes, but in practice such smoothing is unnecessary and even harmful for the conditions examined here. Consistency can be obtained, for example, by only smoothing for sample sizes of over 1,000,000. If smoothing is undesired for some reason in an application or choice of the kernel width seems problematic for the Pruitt estimator, use of the unsmoothed GMLE is recommended. This estimator is somewhat more difficult to compute (see Appendix A.1), but that will be a limit in only a small number of problems where it is not also a limit for the other estimators. Improvements in computing will also push this boundary to higher and higher sample sizes.

# A Appendix.

## A.1 S functions and availability.

The C functions used in the simulations are available as functions which may be run under S. The functions are available by anonymous ftp in the directory */pub/bivsurv* of *umnstat.stat.umn.edu* (128.101.51.1). The C routines are not optimized to save time or memory. For independent exponential random variables, relative computation times are given in Table 3 for a DECstation 3100. Dabrowska and pathwise take a lot of memory to store the answer since they assign mass to so many points, and GML and Pruitt use large intermediate arrays in the EM algorithm computations.

| | Sample size | | | | |
|---|---|---|---|---|---|
| | 50 | 100 | 200 | 400 | 800 |
| Pruitt | 1 | 3 | 7 | 24 | dnf |
| Dabrowska | 2 | 2 | 6 | dnf | dnf |
| GML | 6 | 12 | dnf | dnf | dnf |
| TLC | 1 | 1 | 1 | 4 | 7 |
| Pathwise | 2 | 3 | 9 | dnf | dnf |

Table 3: Time taken to compute the estimate for independent exponential random variables and various sample sizes. dnf shows that the estimator was unable to be computed due to memory limitations.

## A.2 GMLE starting point for the EM algorithm.

The estimator actually used for the GML was a smoothed version of the GML described in Pruitt (1991b). This estimator depends on a smoothing constant (window width), and some discussion of the choice of smoothing constant is given in Section 5.2. The smoothing constant used in the study was 0.0001, which will give a version of the GML as described by Muñoz (1980) unless two data points have coordinates closer than this value. A starting point for the EM algorithm is still necessary. The starting point is described as follows. Consider the observations as sets in the $(x_1, x_2)$ plane, an uncensored point is $\{x_1\} \times \{x_2\}$, a point with first variable censored is $(x_1, \infty) \times \{x_2\}$, a point with second variable censored is $\{x_1\} \times (x_2, \infty)$, and a doubly censored point is $(x_1, \infty) \times (x_2, \infty)$. Form a partition of $\mathbb{R}_+^2$ from the sets associated with each observation and $\mathbb{R}_+^2$. Let the sets of the partition be $B_j, j = 1, \ldots, m$. Let $A_i$ be the set associated with observation $i$, and let $C_j$ be the number of events $A_i, i = 1, \ldots, n$ in which $B_j$ appears. Let

$$D_i = \max_{j: B_j \in A_i} C_j$$

be the largest $C_j$ for any of the events $B_j$ comprising $A_i$. Now spread mass $1/n$ equally over all the sets $B_j$ comprising $A_i$ such that either 1) $C_j = D_i$ or 2) $B_j$ is an uncensored observation. Do this for every observation. This is the starting point for the EM algorithm.

We illustrate the computations on an example. Suppose the data are given in Table 4 as illustrated in Figure 1. The sets of the partition are indicated in Table 5.

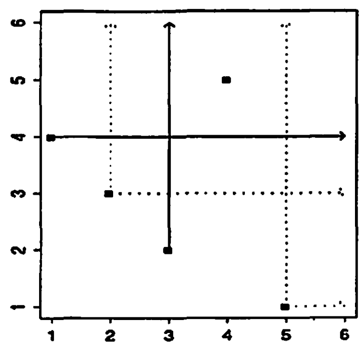


Figure 1: Graphical representation of the data for the example. Observed points are indicated with solid boxes, single censoring with solid arrows, and double censoring by pairs of dotted arrows.

| $i$ | $(Y_1, Y_2, D_1, D_2)$ | $A_i$ | $B_j$ composition | $D_i$ | Initial mass |
|---|---|---|---|---|---|
| 1 | (1,4,0,1) | $(1,\infty) \times \{4\}$ | 2,3,4,8 | 3 | $\frac{1}{8}$ to $B_2, B_3$ |
| 2 | (2,3,0,0) | $(2,\infty) \times (3,\infty)$ | 1,2,3,4,5,6,7 | 3 | $\frac{1}{12}$ to $B_1, B_2, B_3$ |
| 3 | (3,2,1,0) | $\{3\} \times (2,\infty)$ | 2,5,9 | 3 | $\frac{1}{4}$ to $B_2$ |
| 4 | (4,5,1,1) | $\{4\} \times \{5\}$ | 1 | 2 | $\frac{1}{4}$ to $B_1$ |
| 5 | (5,1,0,0) | $(5,\infty) \times (1,\infty)$ | 3,6,10 | 3 | $\frac{1}{4}$ to $B_3$ |

Table 4: The data and decomposition of the associated sets for the example. The $B_j$ decomposition refers to the indices of the $B_j$ given in Table 5.

| $j$ | $B_j$ | $C_j$ |
|---|---|---|
| 1 | $A_2A_4$ | 2 |
| 2 | $A_1A_2A_3$ | 3 |
| 3 | $A_1A_2A_5$ | 3 |
| 4 | $A_1A_2 \setminus (B_2 \cup B_3)$ | 2 |
| 5 | $A_2A_3 \setminus B_2$ | 2 |
| 6 | $A_2A_5 \setminus B_3$ | 2 |
| 7 | $A_2 \setminus (\cup_{j=1}^{6} B_j)$ | 1 |
| 8 | $A_1 \setminus (\cup_{j=2}^{4} B_j)$ | 1 |
| 9 | $A_3 \setminus (B_2 \cup B_5)$ | 1 |
| 10 | $A_5 \setminus (B_3 \cup B_6)$ | 1 |
| 11 | $\mathbb{R}_+^2 \setminus (\cup_{j=1}^{10} B_j)$ | 0 |

Table 5: The partition elements for the example.

## A.3 Integrated squared error formula.

In this subsection we derive the form for the integrated squared error for a step function estimator $\bar{\hat{F}}$ in the Marshall-Olkin model. We actually describe the computations for the more general Marshall-Olkin model given by

$$\bar{F}(x_1, x_2) = \exp\{-\lambda_1 x_1 - \lambda_2 x_2 - \Lambda(x_1 \vee x_2)\}.$$

In our simulations we take $\lambda_1 = \lambda_2 \equiv \lambda$ since 4 of the estimators are scale equivariant. Suppose we have an estimator which is concentrated on points $(s_i, t_j)$ for $i = 0, 1, \ldots, m; j = 0, 1, \ldots, n$ and $0 = s_0 < s_1 < \cdots < s_m; 0 = t_0 < t_1 < \cdots < t_n$. Without loss of generality assume $s_m \geq t_n$. We can then write the integrated squared error for $\bar{\hat{F}}$, (4.1), as

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\int_{s_{i-1}}^{s_i}(\bar{\hat{F}}(s,t) - \bar{F}(s,t))^2 ds dt + \int_0^{t_n}\int_{s_m}^{\infty}\bar{F}(s,t)^2 ds dt$$
$$+ \int_{t_n}^{\infty}\int_0^{t_n}\bar{F}(s,t)^2 ds dt + \int_{t_n}^{\infty}\int_{t_n}^{\infty}\bar{F}(s,t)^2 ds dt. \qquad \text{(A.1)}$$

To ease the writing of solutions define

$$f(\lambda, u) = \frac{e^{-\lambda u}}{\lambda} \qquad \text{and} \qquad d(\lambda, u_1, u_2) = \frac{e^{-\lambda u_1} - e^{-\lambda u_2}}{\lambda}. \qquad \text{(A.2)}$$

We can readily compute the final three integrals to find

$$\int_0^{t_n}\int_{s_m}^{\infty}\bar{F}(s,t)^2 ds dt = f(2(\lambda_1 + \Lambda), s_m)\, d(2\lambda_2, 0, t_n),$$

| Case | Conditions | | | |
|---|---|---|---|---|
| 1 | $s_2 \geq t_1$ | $s_1 \geq t_1$ | $s_2 \geq t_2$ | $s_1 \geq t_2$ |
| 2 | $s_2 \geq t_1$ | $s_1 \geq t_1$ | $s_2 \geq t_2$ | $s_1 < t_2$ |
| 3 | $s_2 \geq t_1$ | $s_1 < t_1$ | $s_2 \geq t_2$ | $s_1 < t_2$ |
| 4 | $s_2 \geq t_1$ | $s_1 \geq t_1$ | $s_2 < t_2$ | $s_1 < t_2$ |
| 5 | $s_2 \geq t_1$ | $s_1 < t_1$ | $s_2 < t_2$ | $s_1 < t_2$ |
| 6 | $s_2 < t_1$ | $s_1 < t_1$ | $s_2 < t_2$ | $s_1 < t_2$ |

Table 6: Indices for the six types of finite boxes.

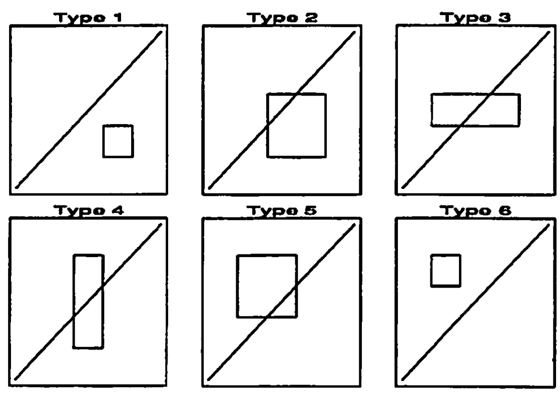


Figure 2: Examples of the six types of finite boxes.

$$\int_{t_n}^{\infty}\int_{0}^{t_n} \bar{F}(s,t)^2 ds dt = f(2(\lambda_2+\Lambda),t_n)\, d(2\lambda_1,0,t_n), \tag{A.3}$$

$$\text{and} \int_{t_n}^{\infty}\int_{t_n}^{\infty} \bar{F}(s,t)^2 ds dt = \left[\frac{1}{2(\lambda_1+\Lambda)}+\frac{1}{2(\lambda_2+\Lambda)}\right] f(2(\lambda_1+\lambda_2+\Lambda),t_n).$$

Suppose we wish to compute

$$I(s_1,t_1,s_2,t_2) = \int_{t_1}^{t_2}\int_{s_1}^{s_2} (\hat{\bar{F}}(s,t)-\bar{F}(s,t))^2 ds dt. \tag{A.4}$$

We separate into six cases according to Table 6. Representative pictures of these types of regions are drawn in Figure 2. As can be seen, Types 4,5, and 6 are reflections of Types 3,2, and 1. We only display formulas for Types 1,2, and 3, as the others may be obtained by reflection.

Let $\vec{\lambda} = (\lambda_1,\lambda_2,\Lambda)$, and note that $\hat{\bar{F}}(s,t) = \hat{\bar{F}}(s_1,t_1)$ for every $s$ and $t$ such that $s_1 \leq s < s_2, t_1 \leq t < t_2$. For Type 1,

$$I(s_1,t_1,s_2,t_2) = (t_2-t_1)(s_2-s_1)\hat{\bar{F}}(s_1,t_1)^2 - 2\hat{\bar{F}}(s_1,t_1)H_1(\vec{\lambda}) + H_1(2\vec{\lambda}), \tag{A.5}$$

where

$$H_1(\vec{\lambda}) = d(\lambda_1 + \Lambda, s_1, s_2)\, d(\lambda_2, t_1, t_2). \tag{A.6}$$

For Type 3, by integrating $s$ first,

$$I(s_1, t_1, s_2, t_2) = (t_2 - t_1)(s_2 - s_1)\hat{\bar{F}}(s_1, t_1)^2 - 2\hat{\bar{F}}(s_1, t_1)H_3(\vec{\lambda}) + H_3(2\vec{\lambda}), \tag{A.7}$$

where

$$\begin{aligned} H_3(\vec{\lambda}) &= f(\lambda_1, s_1)\, d(\lambda_2 + \Lambda, t_1, t_2) - \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_1 + \Lambda}\right) d(\lambda_1 + \lambda_2 + \Lambda, t_1, t_2) \\ &\quad - f(\lambda_1 + \Lambda, s_2)\, d(\lambda_2, t_1, t_2). \end{aligned} \tag{A.8}$$

For Type 2, by writing the integral as a sum of a Type 1 and Type 3 integral,

$$I(s_1, t_1, s_2, t_2) = (t_2 - t_1)(s_2 - s_1)\hat{\bar{F}}(s_1, t_1)^2 - 2\hat{\bar{F}}(s_1, t_1)H_2(\vec{\lambda}) + H_2(2\vec{\lambda}), \tag{A.9}$$

where

$$\begin{aligned} H_2(\vec{\lambda}) &= d(\lambda_1 + \Lambda, s_1, s_2)\, d(\lambda_2, t_1, s_1) + f(\lambda_1, s_1)\, d(\lambda_2 + \Lambda, s_1, t_2) \\ &\quad - \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_1 + \Lambda}\right) d(\lambda_1 + \lambda_2 + \Lambda, s_1, t_2) - f(\lambda_1 + \Lambda, s_2)\, d(\lambda_2, s_1, t_2). \end{aligned} \tag{A.10}$$

Combining (A.1) - (A.10) gives us a formula for the integrated squared error of a step function estimator in the Marshall-Olkin model.

## References


CAMPBELL, G. (1981). Nonparametric bivariate estimation with randomly censored data. *Biometrika* **68** 417-422.

CAMPBELL, G., and FÖLDES, A. (1982). Large sample properties of nonparametric bivariate estimators with censored data. In *Nonparametric Statistical Inference, Colloquia Mathematica-Societatis János Bolyai* (B.V. Gnedenko, M.L. Puri, and I. Vincze, eds.). North Holland, Amsterdam.

DABROWSKA, D.M. (1988). Kaplan-Meier estimate on the plane. *Ann. Statist.* **16** 1475-1489.

DOWNTON, F. (1970). Bivariate exponential distributions in reliability theory. *J. Roy. Statist. Soc. Ser. B* **32** 408-417.

HANLEY, J.A., and PARNES, M.N. (1983). Nonparametric estimation of a multivariate distribution in the presence of censoring. *Biometrics* **39** 129-139.

LANGBERG, N.A. and SHAKED, M. (1982). On the identifiability of multivariate life distribution functions. *Ann. Probab.* **10** 773-779.

LEURGANS, S., TSAI, W.-Y., and CROWLEY, J. (1982). Freund's bivariate exponential distribution and censoring. In *Survival Analysis* (J. Crowley and R. A. Johnson, eds.) 230-242. IMS, Hayward, Calif.

MUÑOZ, A. (1980). Nonparametric estimation from censored bivariate observations. *Technical Report* **60**, Stanford University.

PRUITT, R. (1990). Identifiability of discrete bivariate survival curves from censored data. *Technical Report* **535**, University of Minnesota.

PRUITT, R. (1991a). Strong consistency of self-consistent estimators: general results and an application to bivariate survival analysis. *Preprint*, University of Minnesota.

PRUITT, R. (1991b). Smoothing the generalized maximum likelihood estimator for bivariate survival data. *In preparation*, University of Minnesota.

TSAI, W.-Y., LEURGANS, S., and CROWLEY, J. (1986). Nonparametric estimation of a bivariate survival function in the presence of censoring. *Ann. Statist.* **14** 1351-1365.

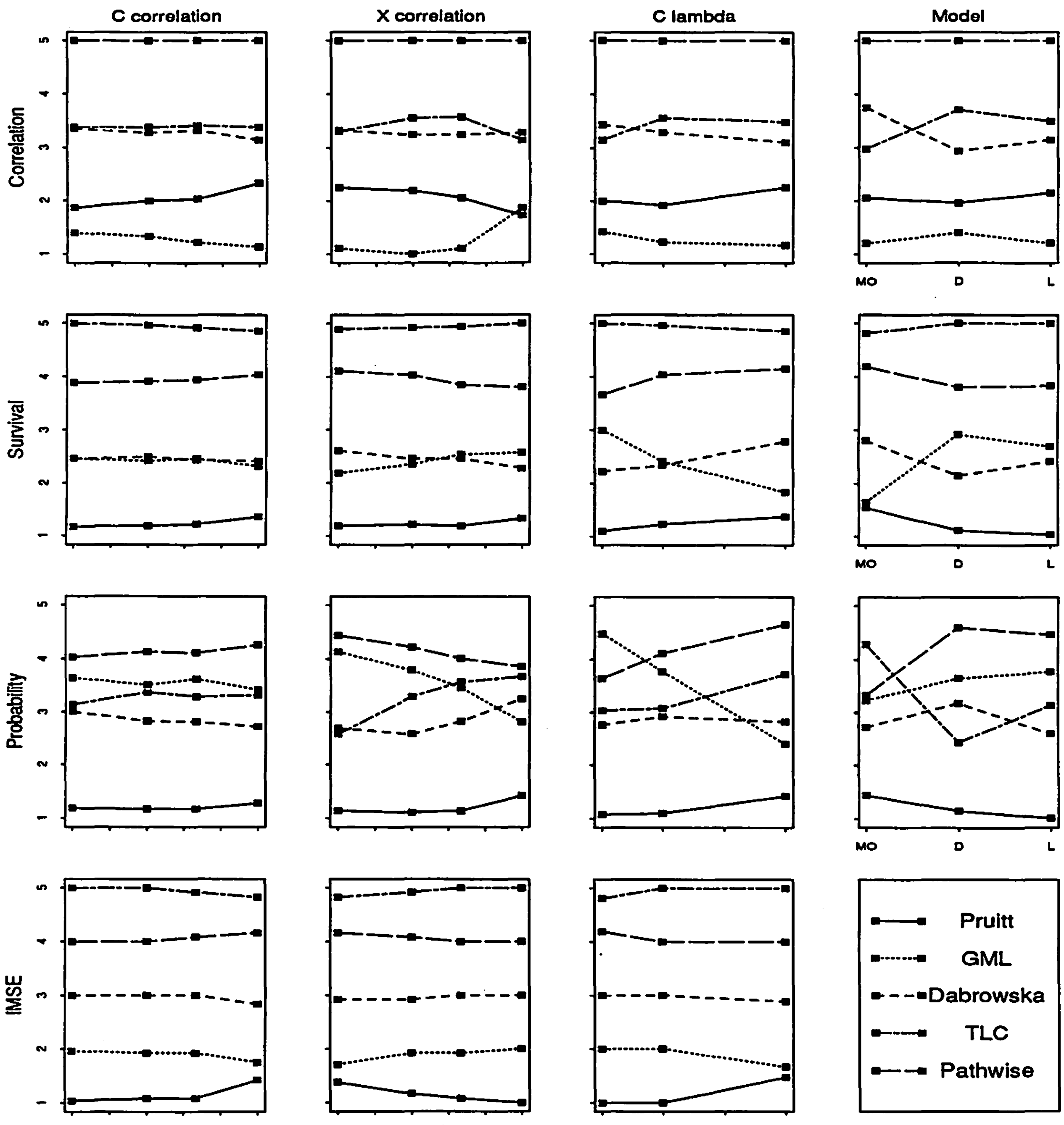


Figure 3: Average relative rank of mean squared error for the estimators. The different problems are across rows and different conditions in columns. The sample size is 50, and the models are MO=Marshall-Olkin, D=Downton, L=linear combination of exponentials. For example, in the upper left graph, the average rank of the GML estimator in the correlation problem over all conditions where the $C$ correlation is zero is 1.39.

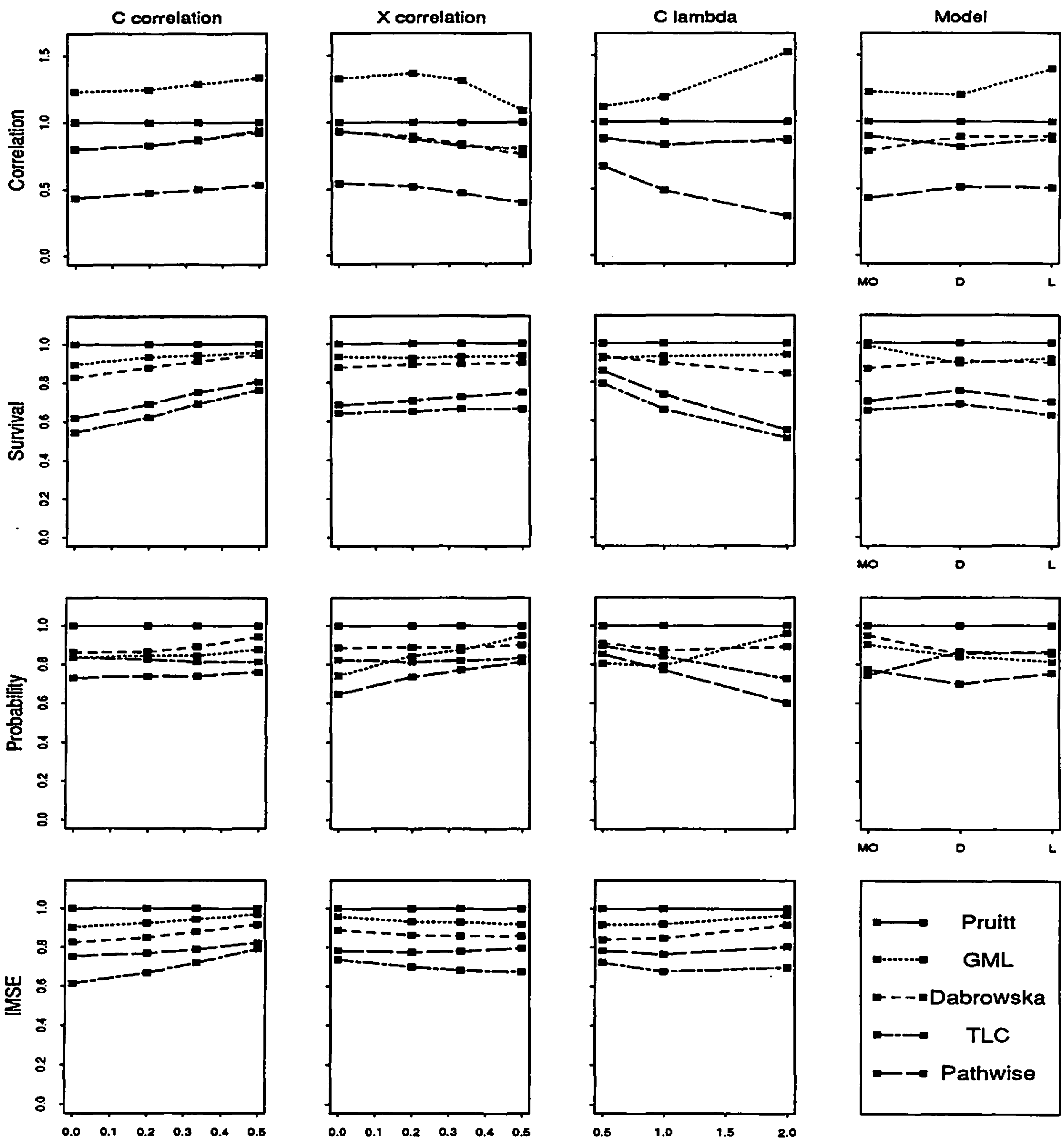


Figure 4: Average relative (to Pruitt) mean squared error for the estimators. The different problems are across rows and different conditions in columns. The sample size is 50, and the models are MO=Marshall-Olkin, D=Downton, L=linear combination of exponentials. For example, in the upper left graph, the average ratio of the MSE for the Pruitt estimate to that of the GML estimate in the correlation problem over all conditions where the $C$ correlation is zero is 1.23.

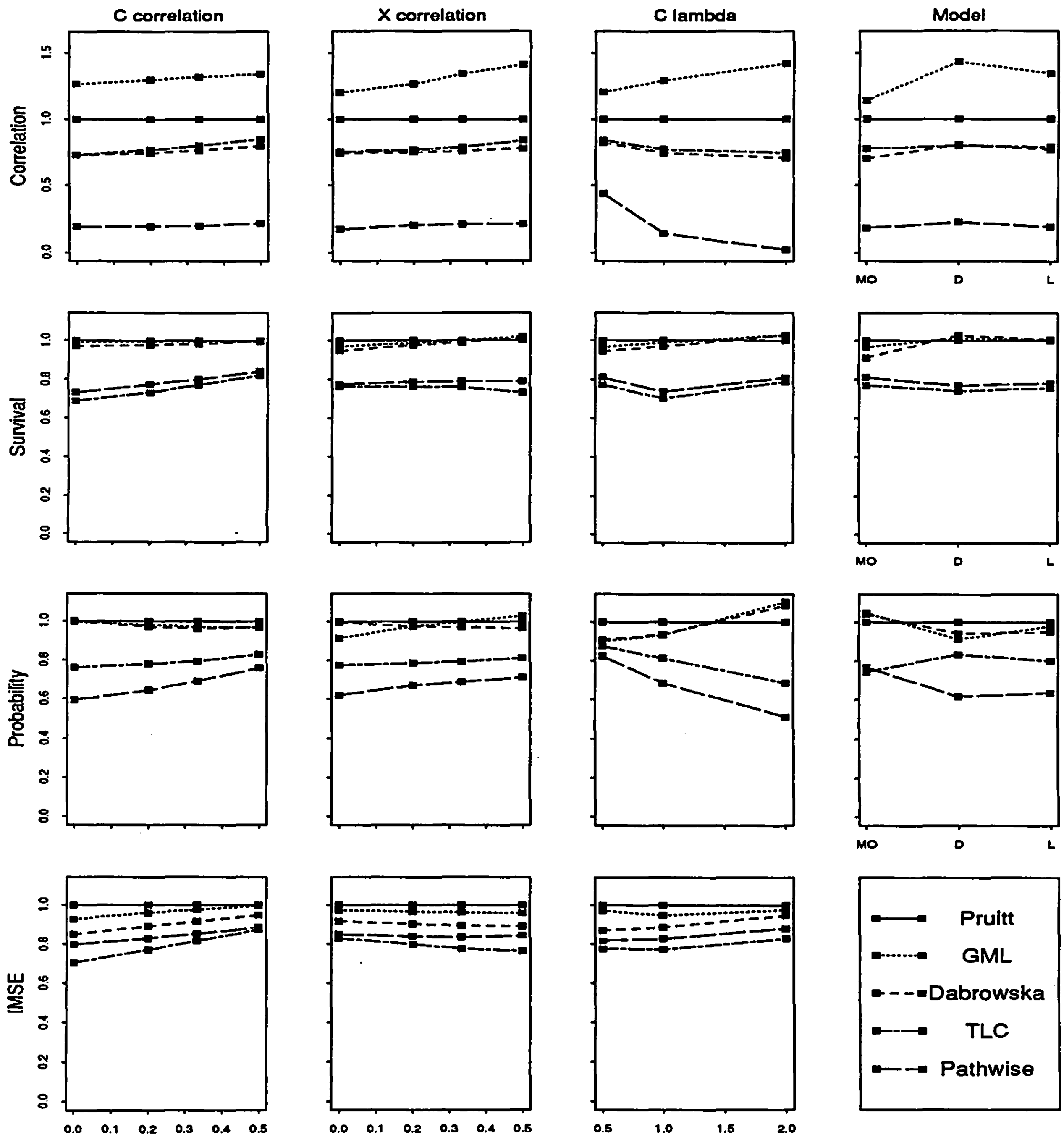


Figure 5: This is Figure 4 for sample size 10. Average relative (to Pruitt) mean squared error for the estimators. The different problems are across rows and different conditions in columns. The sample size is 10, and the models are MO=Marshall-Olkin, D=Downton, L=linear combination of exponentials. For example, in the upper left graph, the average ratio of the MSE for the Pruitt estimate to that of the GML estimate in the correlation problem over all conditions where the $C$ correlation is zero is 1.26.

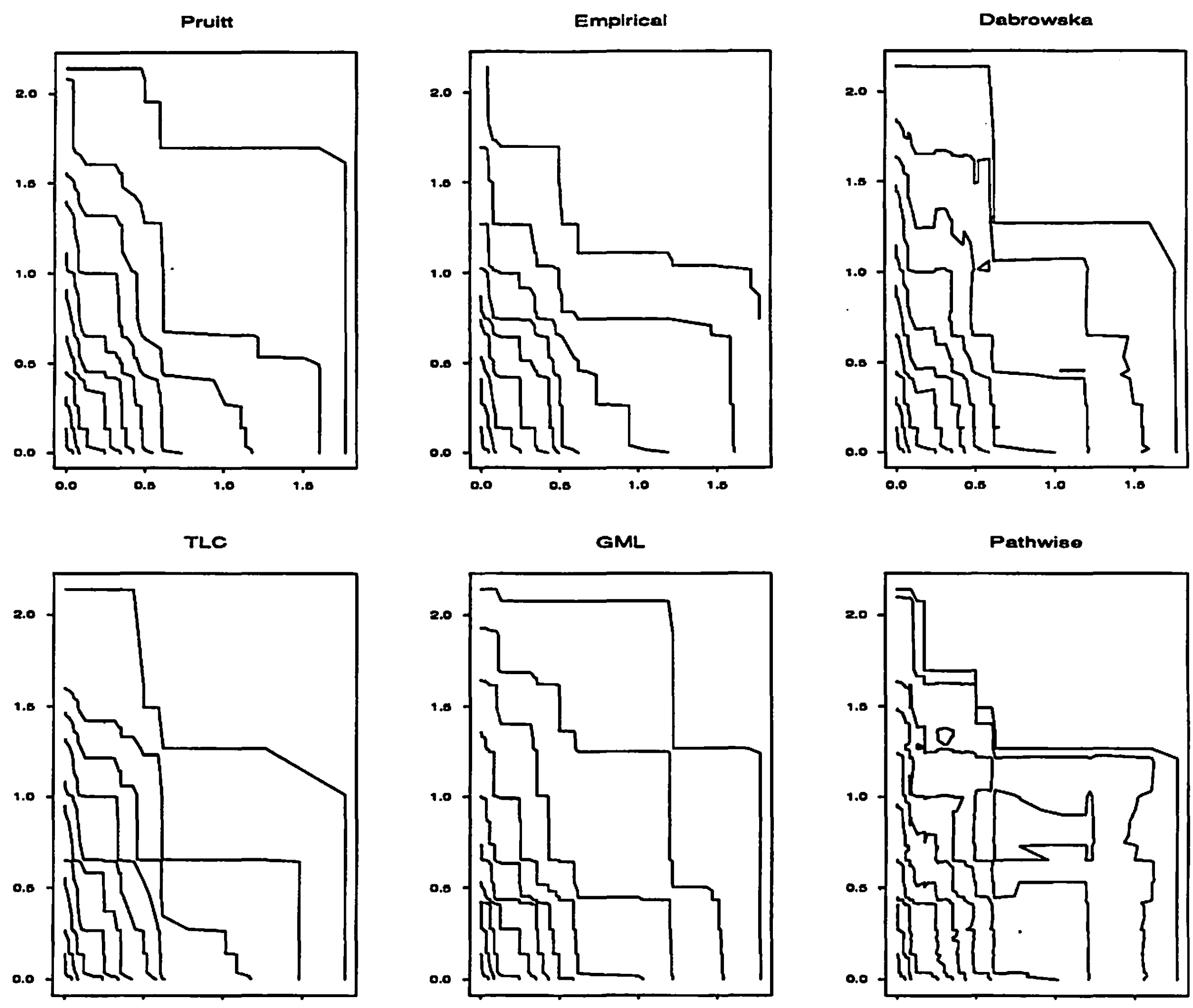


Figure 6: Contour plots for the five estimators and the empirical survival function for a sample of 50 observations with $X_1, X_2, C_1, C_2$ i.i.d. exponential with mean 1.

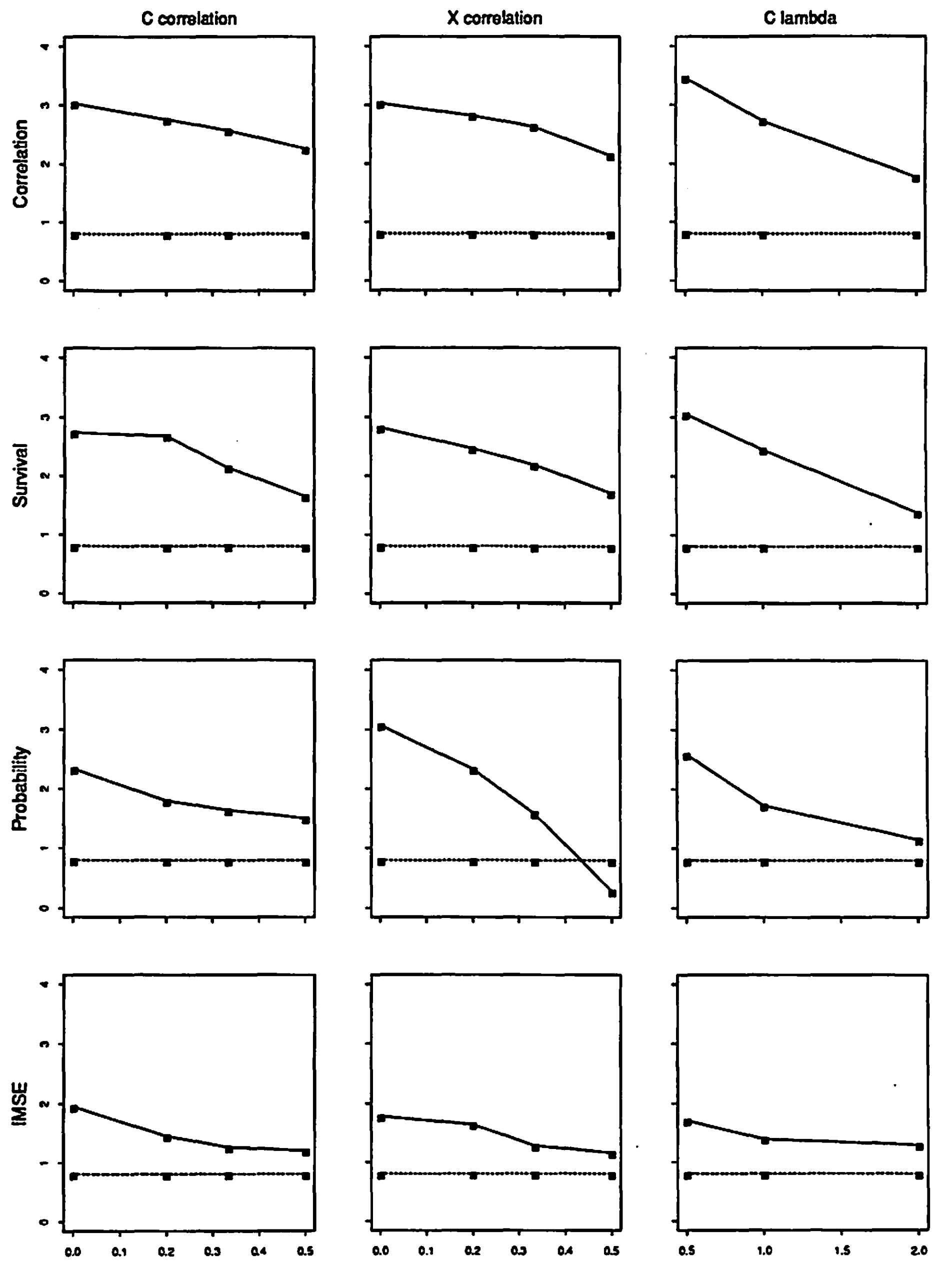


Figure 7: Optimal kernel width for the Pruitt estimator chosen as the smallest kernel from 0.01(0.25)6.01 which achieved minimum mean squared error. Only the conditions for the Marshall-Olkin model were examined.

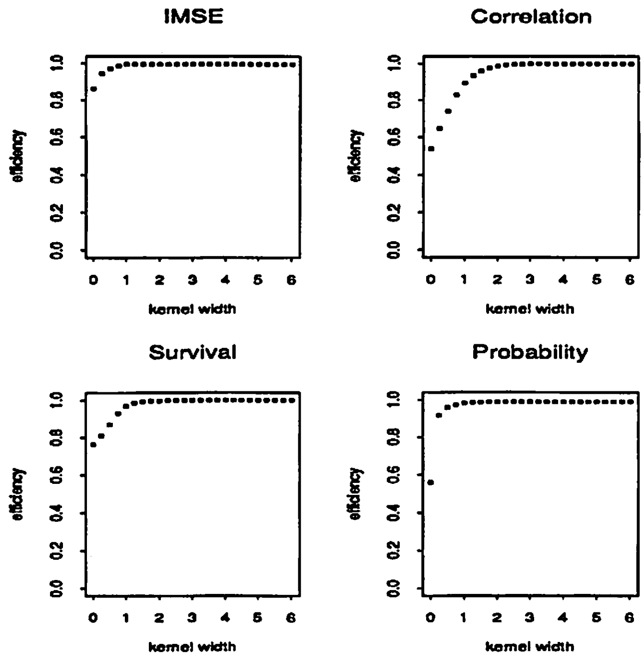


Figure 8: Relative efficiencies of kernel widths 0.01(0.25)6.01 averaged over all conditions for Pruitt estimator in the Marshall-Olkin model.

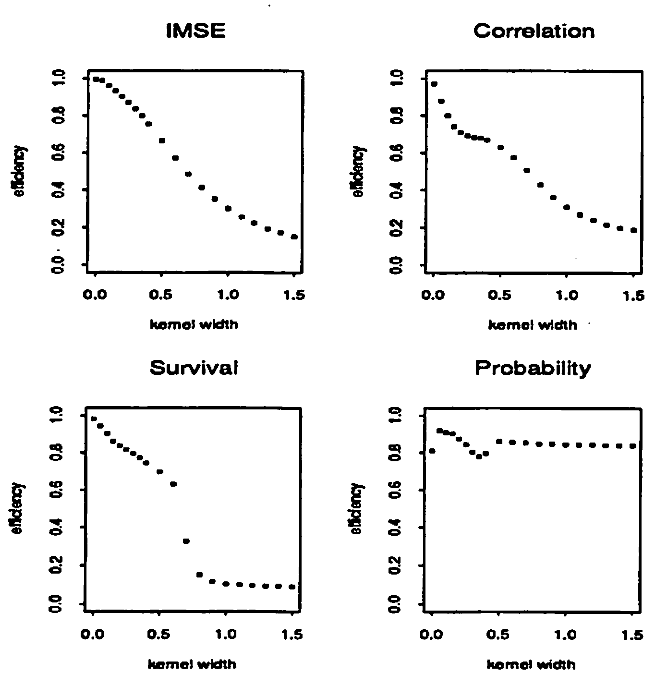


Figure 9: Relative efficiencies of kernel widths 0.0001(0.05)0.4001, 0.5001(0.10)1.5001 averaged over all conditions for GML estimator in the Marshall-Olkin model.

| Estimator | Rank 1 | 2 | 3 | 4 | 5 | Average Rank |
|---|---|---|---|---|---|---|
| Pruitt | .16 | .67 | .13 | .04 | .00 | 2.06 |
| Dabrowska | .00 | .14 | .44 | .42 | .00 | 3.28 |
| GML | .83 | .12 | .01 | .04 | .00 | 1.27 |
| TLC | .01 | .08 | .41 | .50 | .00 | 3.40 |
| Pathwise | .00 | .00 | .00 | .00 | 1.00 | 5.00 |

Table 7: Percentage of cases each estimator had each rank for the correlation problem.

| Estimator | Rank 1 | 2 | 3 | 4 | 5 | Average Rank |
|---|---|---|---|---|---|---|
| Pruitt | .81 | .14 | .05 | .00 | .00 | 1.24 |
| Dabrowska | .05 | .43 | .51 | .00 | .00 | 2.46 |
| GML | .13 | .43 | .32 | .12 | .00 | 2.42 |
| TLC | .00 | .00 | .00 | .06 | .94 | 4.94 |
| Pathwise | .00 | .00 | .12 | .82 | .06 | 3.95 |

Table 8: Percentage of cases each estimator had each rank for the survival problem.

| Estimator | Rank 1 | 2 | 3 | 4 | 5 | Average Rank |
|---|---|---|---|---|---|---|
| Pruitt | .84 | .12 | .04 | .00 | .00 | 1.20 |
| Dabrowska | .01 | .28 | .58 | .13 | .00 | 2.83 |
| GML | .12 | .18 | .08 | .28 | .34 | 3.55 |
| TLC | .00 | .35 | .23 | .20 | .22 | 3.28 |
| Pathwise | .03 | .07 | .07 | .39 | .44 | 4.14 |

Table 9: Percentage of cases each estimator had each rank for the probability problem.

| Problem | | IMSE | | Correlation | | |
|---|---|---|---|---|---|---|
| Estimator | Average size | Average | Relative efficiency | Absolute Bias | MSE | Relative efficiency |
| Pruitt | 135 | 0.0265 | 1.00 | 0.060 | 0.0699 | 1.00 |
| Dabrowska | 576 | 0.0293 | 0.87 | 0.033 | 0.0829 | 0.78 |
| GML | 74.4 | 0.0275 | 0.94 | 0.063 | 0.0525 | 1.22 |
| TLC | 55.2 | 0.0362 | 0.70 | 0.049 | 0.0723 | 0.89 |
| Pathwise | 1033 | 0.0331 | 0.79 | 0.168 | 0.2350 | 0.43 |

| Problem | Survival | | | Probability | | |
|---|---|---|---|---|---|---|
| Estimator | Absolute Bias | MSE | Relative efficiency | Absolute Bias | MSE | Relative efficiency |
| Pruitt | 0.035 | 0.0102 | 1.00 | 0.018 | 0.0106 | 1.00 |
| Dabrowska | 0.024 | 0.0110 | 0.87 | 0.004 | 0.0103 | 0.95 |
| GML | 0.038 | 0.0100 | 0.99 | 0.020 | 0.0106 | 0.90 |
| TLC | 0.040 | 0.0155 | 0.65 | 0.023 | 0.0196 | 0.75 |
| Pathwise | 0.042 | 0.0151 | 0.70 | 0.010 | 0.0161 | 0.78 |

Table 10: Marshall-Olkin model averages. Note the average relative efficiency is the average of the relative efficiencies over the 48 conditions, not the relative efficiency of the average MSE.

| Problem | | Correlation | | |
|---|---|---|---|---|
| Estimator | Average size | Absolute Bias | MSE | Relative efficiency |
| Pruitt | 120 | 0.060 | 0.0781 | 1.00 |
| Dabrowska | 584 | 0.052 | 0.0927 | 0.89 |
| GML | 73.5 | 0.111 | 0.0707 | 1.20 |
| TLC | 55.5 | 0.092 | 0.1040 | 0.82 |
| Pathwise | 1069 | 0.274 | 0.3010 | 0.51 |

| Problem | Survival | | | Probability | | |
|---|---|---|---|---|---|---|
| Estimator | Absolute Bias | MSE | Relative efficiency | Absolute Bias | MSE | Relative efficiency |
| Pruitt | 0.0047 | 0.00675 | 1.00 | 0.00094 | 0.00181 | 1.00 |
| Dabrowska | 0.0013 | 0.00781 | 0.91 | 0.00065 | 0.00214 | 0.86 |
| GML | 0.0212 | 0.00765 | 0.89 | 0.00797 | 0.00225 | 0.84 |
| TLC | 0.0077 | 0.01370 | 0.68 | 0.00070 | 0.00212 | 0.87 |
| Pathwise | 0.0084 | 0.01250 | 0.75 | 0.00075 | 0.00277 | 0.70 |

Table 11: Downton model averages. Note the average relative efficiency is the average of the relative efficiencies over the 48 conditions, not the relative efficiency of the average MSE.

| Problem | | Correlation | | |
|---|---|---|---|---|
| Estimator | Average size | Absolute Bias | MSE | Relative efficiency |
| Pruitt | 122 | 0.042 | 0.0614 | 1.00 |
| Dabrowska | 567 | 0.021 | 0.0686 | 0.90 |
| GML | 72.0 | 0.051 | 0.0439 | 1.40 |
| TLC | 52.1 | 0.037 | 0.0712 | 0.87 |
| Pathwise | 1046 | 0.173 | 0.1600 | 0.51 |

| Problem | Survival | | | Probability | | |
|---|---|---|---|---|---|---|
| Estimator | Absolute Bias | MSE | Relative efficiency | Absolute Bias | MSE | Relative efficiency |
| Pruitt | 0.0076 | 0.00757 | 1.00 | 0.0031 | 0.00309 | 1.00 |
| Dabrowska | 0.0015 | 0.00873 | 0.90 | 0.0009 | 0.00356 | 0.87 |
| GML | 0.0222 | 0.00824 | 0.92 | 0.0108 | 0.00379 | 0.82 |
| TLC | 0.0120 | 0.01590 | 0.63 | 0.0015 | 0.00361 | 0.86 |
| Pathwise | 0.0132 | 0.01470 | 0.70 | 0.0010 | 0.00422 | 0.76 |

Table 12: Linear model averages. Note the average relative efficiency is the average of the relative efficiencies over the 48 conditions, not the relative efficiency of the average MSE.

| Problem | | Correlation | | |
|---|---|---|---|---|
| Estimator | Average size | Bias | MSE | Relative efficiency |
| Pruitt | 125 | -0.409 | 0.1895 | 1.00 |
| Dabrowska | 279 | -0.017 | 0.2190 | 0.87 |
| GML | 82.8 | -0.593 | 0.4416 | 0.43 |
| TLC | 10.9 | -0.084 | 0.1003 | 1.89 |
| Pathwise | 498 | -0.409· | 1.8101 | 0.10 |

| Problem | Survival | | | Probability | | |
|---|---|---|---|---|---|---|
| Estimator | Bias | MSE | Relative efficiency | Bias | MSE | Relative efficiency |
| Pruitt | -0.0696 | 0.01978 | 1.00 | -0.0117 | 0.00662 | 1.00 |
| Dabrowska | 0.0018 | 0.01925 | 1.03 | -0.0002 | 0.00719 | 0.92 |
| GML | -0.0698 | 0.01871 | 1.06 | -0.0241 | 0.00722 | 0.92 |
| TLC | -0.0061 | 0.03499 | 0.57 | -0.0024 | 0.01407 | 0.47 |
| Pathwise | -0.0104 | 0.02225 | 0.89 | -0.0186 | 0.01321 | 0.50 |

Table 13: Results for sample size 50 and the mixture of normals model.