Biometrics 99, 1–7 DOI: xx.xxxx/x.xxxx-xxxx.xxxx.xxxxx.x
March 2013

Supplementary Web Section:

Semiparametric Bayesian commensurate survival model for post-market medical device surveillance with non-exchangeable historical data

Thomas A. Murray[1,*], Brian P. Hobbs[2], Theodore C. Lystig[3], and Bradley P. Carlin[1]

[1]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, U.S.A.

[2]Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, U.S.A.

[3]Medtronic Inc., Minneapolis, MN, U.S.A.

email: murra484@umn.edu

This paper has been submitted for consideration for publication in *Biometrics*

Supplementary Web Section

Simulation Comparison of the Procedures

To assess the performance of our proposed models, we replicate pairs of datasets containing time to event information for the current and historical populations, each with known hazard functions. Simulations are conducted over an array of hazard function combinations. For each simulation, we assess model fit on every replicate dataset pair, and compare models based on average fit over a large number of replicates. Throughout, the replicate dataset pairs contain possibly right-censored observations with no covariates. For a current dataset of size n, we generate failure times $t_i, i = 1, \ldots, n$, from a survival distribution f. We generate censoring times $c_i, i = 1, \ldots, n$, from a uniform distribution and also impose a maximum follow up time (C) at which point all observations are censored. The dataset is then compiled in the usual manner by creating the observed failure times $y_i = \min\{\mathrm{t_i}, \mathrm{c_i}, \mathrm{C}\}$ and event indicators $\delta_i = \mathrm{I}\{\mathrm{t_i} < \min(\mathrm{c_i}, \mathrm{C})\}, \mathrm{i} = 1, \ldots, \mathrm{n}$. The corresponding failure and censoring times for the historical dataset of size n_0 are generated analogously from distribution f_0, possibly different from f. We use the following distributional conventions, $\mathcal{E}(\zeta)$ denotes an exponential distribution with mean ζ^{-1}, and $\mathcal{W}(\psi, \nu) \propto x^{\psi-1} \exp(-\nu x^{\psi})$ denotes a Weibull distribution.

Our first investigation uses exponential distributions to generate the failure times, conducted by assuming $f \sim \mathcal{E}\{e^{\alpha}\}$ and $f_0 \sim \mathcal{E}(1)$. We randomly draw α from the mixture of $0.5 * U(-\log(4), \log(4))$ and point masses at $-\log(4)$, $-\log(3)$, $-\log(2)$, $-\log(1.5)$, $\log(1)$, $\log(1.5)$, $\log(2)$, $\log(3)$, $\log(4)$, each with probability .5/9. The censoring times are drawn from a $U(0, 2)$ distribution with a maximum follow up of $C = 1$. For the range of α values in the exponential simulation, the current versus historical hazard ratio ranges .25 to 4 and $S(1)$ ranges from .78 to .02, respectively. The left panel of Figure 1 displays the true survival distribution of the current population for select values of α. For our second investigation, we generate failure times from Weibull distributions, such that $f \sim \mathcal{W}\{e^{\alpha}, 1\}$ and $f_0 \sim \mathcal{W}(1, 1)$.

For the Weibull simulation, the current versus historical hazard ratio is no longer constant, rather it varies over the follow up period as follows, $e^{\alpha}t^{e^{\alpha}-1}$. For $\alpha > 0$, the hazard ratio decreases over time, and for $\alpha < 0$ it increases over time. The larger in absolute value α is, the more extreme the change in the hazard ratio over the follow up period, but for all α, $S(1) = 0.37$. We again draw α from the same mixture distribution and censoring times from $\mathcal{U}(0, 2)$ with $C = 1$. The right panel of Figure 1 displays the true survival distribution of the current population for select values of α. For both investigations, a larger absolute value of α corresponds to greater heterogeneity in the two sources of data as the true survival distribution of the historical population corresponds to $\alpha = 0$.

[Figure 1 about here.]

Evaluation Criterion

We compare our novel model to the Kaplan-Meier (K-M) estimator fit to both the current data alone and the pooled data. We refer to the former as the "Current K-M" and the latter as the "Pooled K-M". We also fit the fully Bayesian piecewise exponential (PE) model to the current data alone and the pooled data. For both of the fully Bayesian reference models, we use the same time axis partition as our novel model, and employ a random walk prior process on the log-hazards as defined in Equation (3) of the paper. We refer to these models as the "Current PE" and the "Pooled PE". We refer to our novel method as the smoothing commensurate model, or "Comm PE" for short. To assess model performance, we fit each model to each replicate dataset pair generated as detailed above. For the K-M estimators, we save the estimated survival curve and the complementary log-log 95% pointwise confidence interval for $S(.75)$. For the three Bayesian models, we save the posterior mean survival curve and the 95% HPD credible intervals for $S(.75)$. For the sake of notational convenience, we refer only to the posterior survival curve, $S(t|\boldsymbol{D}, \boldsymbol{D_0})$, but in the context of the K-M fits we mean $\widehat{S}(t)$. Using these summaries, we calculate four evaluation criteria, integrated error

(IE), integrated squared error (ISE), interval width, and an indicator for coverage of the true value of $S(.75|\alpha)$.

Formally, for the m^{th} replicate dataset pair, $(\boldsymbol{D}^{(m)}, \boldsymbol{D_0}^{(m)})$, we define and approximate

$$IE^{(m)} = \int_0^{\kappa_K} \left\{ S\left(t|\boldsymbol{D}^{(m)}, \boldsymbol{D_0}^{(m)}\right) - S\left(t|\alpha^{(m)}\right) \right\} dt \tag{1}$$
$$\approx \sum_{g=1}^{G} \nu \left\{ S\left(z_g|\boldsymbol{D}^{(m)}, \boldsymbol{D_0}^{(m)}\right) - S\left(z_g|\alpha^{(m)}\right) \right\},$$

and

$$ISE^{(m)} = \int_0^{\kappa_K} \left\{ S\left(t|\boldsymbol{D}^{(m)}, \boldsymbol{D_0}^{(m)}\right) - S\left(t|\alpha^{(m)}\right) \right\}^2 dt \tag{2}$$
$$\approx \sum_{g=1}^{G} \nu \left\{ S\left(z_g|\boldsymbol{D}^{(m)}, \boldsymbol{D_0}^{(m)}\right) - S\left(z_g|\alpha^{(m)}\right) \right\}^2,$$

$m = 1, \ldots, M$, where M is the total number of replicate dataset pairs generated. The approximations are simple Riemannian ones, calculated by partitioning the time axis into G equally spaced grid points, $0 < z_1 < z_2 < \ldots < z_G = \kappa_K$ where $\nu = \frac{\kappa_K}{G}$ is the mesh of the grid.

Armed with M realizations of IE, ISE, credible interval width at $S(.75)$, and indicators for coverage of the true $S(.75)$ for a range of α values, we can now calculate the expected value for each given α and $\alpha_0 = 1$. To estimate these expected values, we fit a generalized additive model for each of the evaluation criteria and each of the models being compared. For IE, ISE, and credible interval width, we model $E(\text{Criterion}|\alpha) = \text{f}(\alpha)$, and estimate $f(\alpha)$ with a natural cubic spline containing 20 knots equally spaced over the range of α values. Estimation and the choice of the smoothing penalty is conducted by generalized cross validation in the `R` package `mgcv` (Wood, 2006). Similarly, we find the expected coverage probability given α using the same technique, but with a logistic generalized additive model instead. Note that these evaluation criteria are classical frequentist measures of fit; they are estimates of expectations with respect to the sampling distribution, not the posterior distribution.

A final issue ever-present in modern Bayesian work is MCMC convergence monitoring.

We ran a preliminary investigation of convergence for the piecewise exponential models and use 200 iterations of burn-in with 20,000 posterior draws for estimation of the reference models, and 2,000 iterations of burn-in with 20,000 draws for estimation of the commensurate models. The Bayesian models are fit to each replicate dataset pair by calling `JAGS` (Plummer, 2003) through `R` (R Development Core Team, 2011) via the `R2jags` package. Fitting the five methods to each replicate dataset pair and calculating the evaluation statistics takes about 30 seconds, but because of the computational burden of iteratively calling an MCMC sampler like `JAGS`, we also employed the `R` package `snowfall` to run the simulations in parallel, making this investigation computationally feasible.

Results

For both investigations we set the number of replicate dataset pairs to $M = 2,000$, and used a $G = 2,000$ point grid to estimate (1) and (2). Each current dataset had $n = 80$ observations and each historical dataset had $n_0 = 200$ observations. Preliminarily, we investigated an array of specifications for the spike and slab hyperprior on the inter-source smoothing parameter τ in the smoothing commensurate prior model. $\mathcal{S}_u$, $\mathcal{R}$, and p_0 jointly control borrowing of strength based on evidence of heterogeneity in the data. Given $\mathcal{R}$ and p_0, a smaller $\mathcal{S}_u$ imposes greater borrowing as more evidence for heterogeneity in the data is needed to prevent borrowing. Increasing p_0 also results in borrowing more readily. Whereas, given $\mathcal{S}_u$ and p_0, when insufficient evidence for heterogeneity exist in the data, a larger $\mathcal{R}$ results in greater inter-source smoothing. The specification we choose to use for the first simulation is $\mathcal{S}_l = 0.0001$, $\mathcal{S}_u = 2$, $\mathcal{R} = 500$, and $p_0 = 0.99$. This combination of hyperparameter values represents a very large amount of skepticism regarding exchangeability of the two sources of information, with a prior probability of just 0.01 for strong borrowing. For the second simulation we set $\mathcal{S}_l = 0.0001$, $\mathcal{S}_u = 2$, $\mathcal{R} = 200$, and $p_0 = 0.9$ representing less skepticism

regarding the exchangeability of the two sources, but borrowing a bit less strength when the inter-study smoothing parameter falls in the spike.

The results of the first simulation for the K-M and PE models are illustrated in Figure 2. The IME plot (top left panel) evaluates bias by showing how $E(IE|\alpha)$ changes across α, with values near zero indicating an unbiased method for a given α. The IMSE plot (top right panel) evaluates efficiency by showing how $E(ISE|\alpha)$ changes with α, with smaller values indicating a more efficient method for a given α. The credible interval width plot (bottom left panel) is a second measure of efficiency, showing the expected 95% HPD credible interval width for a given α. Finally, the coverage probability plot (bottom right panel) shows the probability that the credible interval will contain the true $S(.75|\alpha)$ for a given α. As evidenced by the IME and coverage probability plots, the current K-M estimator is unbiased and the complementary log-log pointwise confidence interval provides approximate 95% coverage for all α. The current PE model has minor positive bias and a similar coverage profile as the current K-M. Both the current K-M and current PE models have similar efficiency profiles, with the PE model dominating the K-M estimator in both IMSE and interval width. The pooled PE and pooled K-M have very similar profiles for all the evaluation criteria. Namely, both are highly efficient and unbiased when the two sources are truly commensurate (i.e., $\alpha = 0$), but suffer from great bias and poor efficiency as between source heterogeneity increases (i.e., α increases in absolute value). Our novel method shows an intermediate profile to the current and pooled options for all the evaluation criteria. Our method results in efficiency and bias similar to the pooled PE when the two sources are commensurate (i.e., $\alpha \approx 0$) and eventually results in efficiency and bias similar to the current PE when between source heterogeneity is great. We note that in simulations not shown here, adjusting the spike and slab hyperparameter values results in different bias-variance tradeoffs. For instance, increasing the value of p_0 will

result in a more skeptical spike and slab prior and thus, a reduction in the magnitude of bias and efficiency given intermediate between source heterogeneity (i.e., $|\alpha| \approx .5$).

[Figure 2 about here.]

Figure 3 contains the results of the second investigation, where the current and historical populations have crossing Weibull hazard curves with the same true $S(1)$ value. As in the exponential simulations, as the absolute value of α is further from 0, the two survival curves are more dissimilar and the current hazard function changes more sharply across time. The current K-M estimator again shows consistent nearly zero bias and good nominal coverage, but poor efficiency when the two sources are commensurate. The pooled K-M again shows poor bias and efficiency as source heterogeneity grows. The smoothing commensurate PE model (Comm PE) again closely mirrors the current PE model and imposes far less bias as the two hazard curves become sufficiently dissimilar, while providing increased efficiency like the pooled PE model when the two populations are commensurate. The Comm PE model shows better nominal coverage probabilities on average than in the previous exponential investigation. Here, as α departs from 0, the random walk prior process employed in the PE models is no longer smoothing toward the correct model, the result is increasingly poor bias and efficiency for all the PE models. The PE models can be improved greatly by facilitating intra-source smoothing of α_k toward $\rho\alpha_{k-1}$ as in (Fahrmeir and Lang, 2001).

[Figure 3 about here.]

Depending upon the setting, the bias-efficiency profile of the smoothing commensurate model may not seem appropriate. The advantage of this method is that the investigator can control the bias-variance profile through the specification of the spike and slab prior distribution on the inter-source smoothing parameter τ. In simulations not shown here, the bias-efficiency profile of our method is most sensitive to p_0, less sensitive to $\mathcal{R}$ and $\mathcal{S}_u$, and insensitive to $\mathcal{S}_l$. In practice, a variety of combinations should be investigated and a

reasonable bias-efficiency profile should be chosen based on the setting. Thinking of the IME and IMSE curves as sine waves, increasing p_0 will reduce the amplitude and period. Increasing $\mathcal{S}_u$ will also reduce the amplitude and period, but will eventually result in greater bias overall. Finally, increasing $\mathcal{R}$ will reduce the amplitude and have little effect on the period of the IME and IMSE waves, but the amplitude of coverage probability wave will increase. Calibration of this prior should be done with care, and we do not recommend using one particular spike and slab specification for every setting.

References

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50,** 201–220.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Received October 2012.

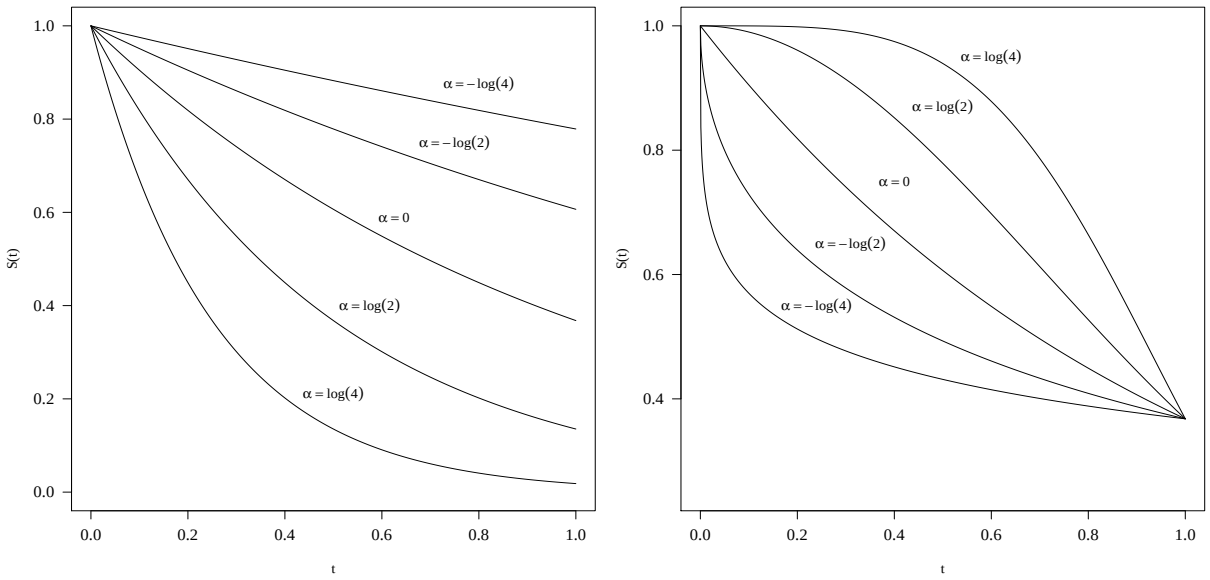

Figure 1. True survival distributions of the current data for selected values of α for the exponential (left panel) and Weibull (right panel) investigations. The true survival distribution for the historical data in both investigations corresponds to $\alpha = 0$.

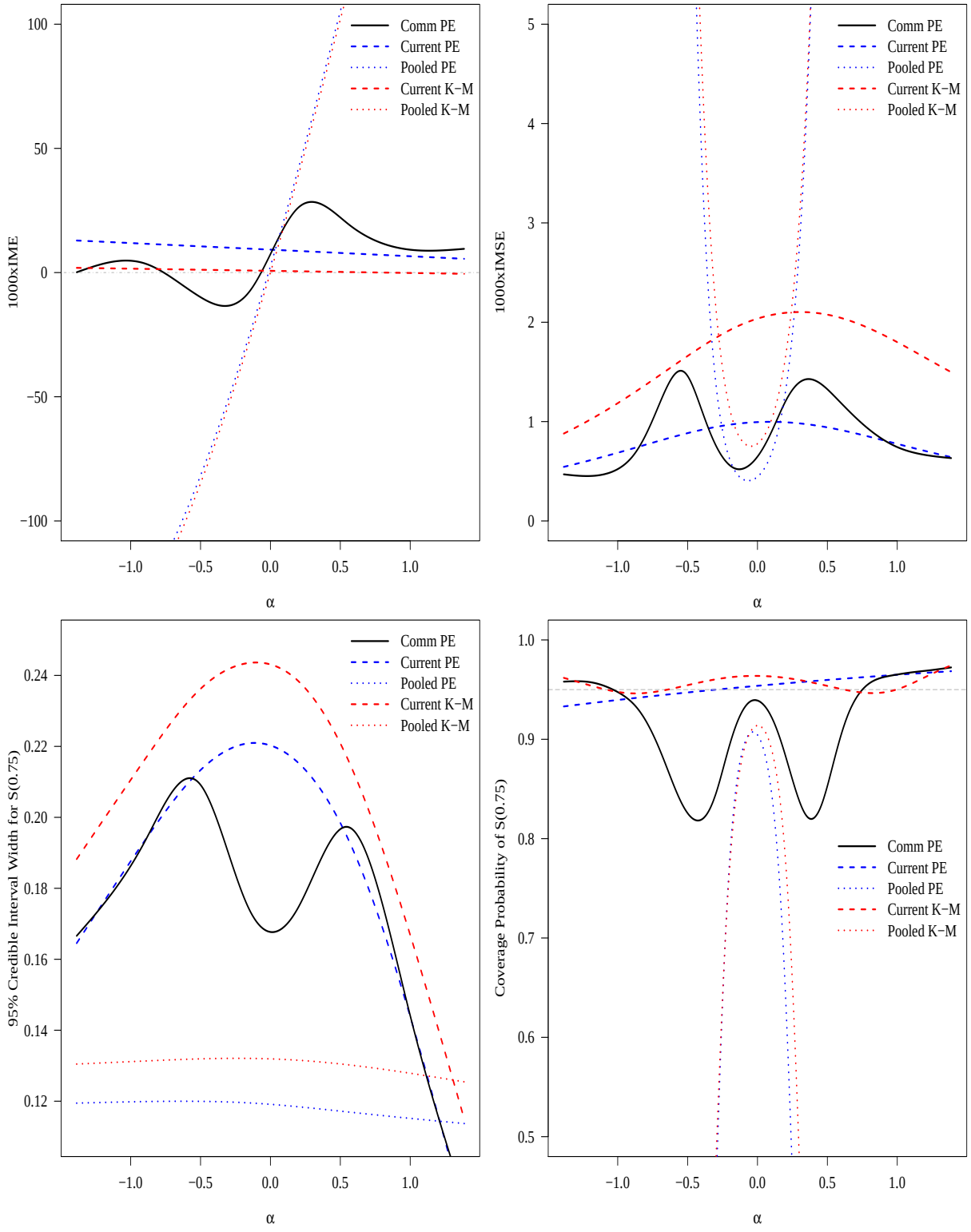

Figure 2. Evaluation criteria of various models fit to dataset pairs generated from exponential distributions. The evaluation criteria are integrated mean error (IME, top left panel), integrated mean squared error (IMSE, top right panel), credible interval width of $S(0.75|\boldsymbol{D}, \boldsymbol{D_0})$ (bottom left panel), and coverage probability of $S(0.75|\alpha)$ (bottom right panel)

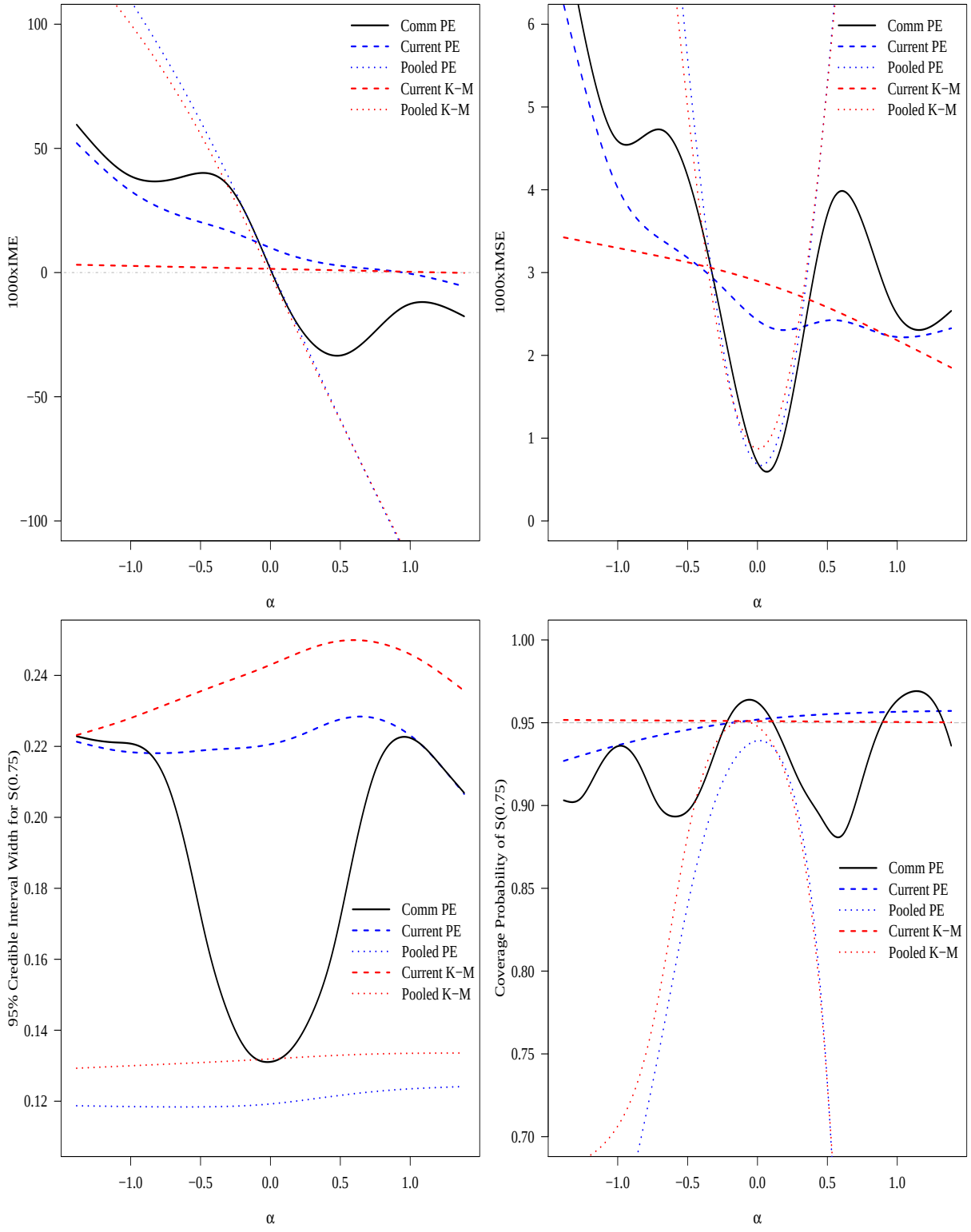

Figure 3. Evaluation criteria of various models fit to dataset pairs generated from Weibull distributions. The evaluation criteria are integrated mean error (IME, top left panel), integrated mean squared error (IMSE, top right panel), credible interval width of $S(0.75|\boldsymbol{D}, \boldsymbol{D_0})$ (bottom left panel), and coverage probability of $S(0.75|\alpha)$ (bottom right panel)