Hi Dr. Agustin.

I've been thinking about what you've asked me to do regarding how to simulate the candidate sets and what the likelihood function is, and I don't think I'll be able to defend that perspective in my paper/presentation. Maybe I'm still missing something.

I believe I found an interesting paper that sheds light on the source of our disagreement. In [1], they write (I have replaced their notation with mine, e.g., $C_i = S_i$, $R_j(t_i) = \bar{F}_j(t_i)$, and so on to make it easier to compare to my work):

> Consider (t_i, C_i) and its contribution l_i to full likelihood L,
>
> $$l_i = f_{T_i,C_i}(t_i, c_i) = \sum_{j \in c_i} \left(f_j(t_i) \prod_{\substack{s=1 \\ s \neq j}}^{m} R_s(t_i) f_{C_i|T_i,K_i}(c_i|t_i, j) \right).$$

where they explain that

> The term $f_{T_i,K_i}(t_i, j) = f_j(t_i) \prod_{s=1, s\neq j}^{m} R_s(t_i)$ is derived from the system failing at time t_i due to cause (i.e., component) j. The expression $f_{C_i|T_i,K_i}(c_i|t_i, c_i)$ represents the conditional probability that the observed minimum random subset c_i *given* that system i failed at time t_i and the true cause was component j.

They clarify that

> In our industrial problems, we found that masking generally occured due to constraints of time and the expense of failure analysis. Schedules often dictated that complete failure analysis be curtailed. In this setting, we found that for fixed j' and in c_i that
>
> $$f_{C_i|T_i,K_i}(c_i|t_i, j') = f_{C_i|T_i,K_i}(c_i|t_i, j) \text{ for all } j \in c_i. \tag{2.2}$$
>
> As a result, this term can be factored out of the summation to yield
>
> $$l_i = f_{T_i,C_i}(t_i, c_i) = f_{C_i|T_i,K_i}(c_i|t_i, j') \sum_{j \in c_i} \left(f_j(t_i) \prod_{\substack{s=1 \\ s \neq j}}^{m} R_s(t_i) \right).$$
>
> The full likelihood under (2.2) is then
>
> $$L = \prod_{i=1}^{n} l_i.$$

They go on to write

> We assume only that masking probabilities conditional on time are not functions of the life distribution parameters. We state this for future reference as:
>
> $$f_{C_i|T_i,K_i}(c_i|t_i, j') \text{ does not depend on the life distribution parameters.} \tag{2.3}$$

A bit later, they explain that

> Using (2.3), we write a reduced or partial likelihood as
>
> $$L_R = \prod_{i=1}^{n} \left\{ \sum_{j \in c_i} f_j(t_i) \prod_{\substack{s=1 \\ s \neq j}}^{m} R_s(t_i) \right\}.$$
>
> Under (2.2) and (2.3), maximizing L_R with respect to the life parameters is equivalent to using L.

They sum this up by saying that it is best to view L_R "as a partial likelihood that under appropriate conditions will yield good, consistent estimators. Without (2.2), and all else true, for example, such a partial likelihood can yield inconsistent estimators." Finally, they write: "For proper statistical applications, we stress that it is important to be clearly aware of the effects of masking probabilities and needed conditions."

So, this is the source of the confusion. We are not assuming condition (2.2) holds, i.e., we do not assume that

$$f_{C_i|T_i,K_i}(c_i|t_i,j') = f_{C_i|T_i,K_i}(c_i|t_i,j) \text{ for all } j \in c_i..$$

It is very unfortunate that many papers on this topic did not explain their assumptions very well. I had to look pretty hard to find a paper that finally articulated these assumptions for when maximizing L_R produces the same estimates as maximizing L.

So, now let's turn back to the approach in my draft. In my statistical model, we assume that the failed component is in the candidate set. This is often a real-world situation, e.g., replacing a circuit board with multiple components fixes the system, so we only know that one of those components on the circuit board was the cause of failure. We do not assume condition (2.2), since our Bernoulli candidate model is explicitly allowed to violate it in most cases.

However, the way you have described simulating the data, candidate sets and the series system do not have any statistical correlations and can therefore the simulated candidate sets cannot provide any information about the system. Moreover, the likelihood function L_R is inappropriate for our simulated data, which explicitly is allowed to violate condition (2.2).

So, there is some underlying, unknown data generating process for candidate sets. Suppose the true joint pdf of the failure time and the candidate set is given by

$$f^*_{T_i,C_i}(t_i,c_i)$$

for $i = 1,\ldots,n$.

We do not know this joint distribution, so we replace $f^*_{T_i,C_i}(t_i,c_i)$ with some parametric pdf that, with an appropriate parameterization, will hopefully be able to reasonably approximate the true data generating process. (All models are wrong, some are useful.) This is no different than using a Weibull distribution to model lifetime data in a real-world data set, no different than approximating a complex system with a series system model with independently distributed component lifetimes, and no different than assuming an i.i.d. random sample. We use simplifying assumptions about the data generating process to make the problem tractable, explainable, and understandable.

So, we do not know the true data generating process, but we look at the data and try to choose a reasonable parametric family that we believe, with an appropriate choice of parameter values (using MLE, in our case), will provide a reasonable approximation of the true data generating process.

So, I am seeking to approximate $f^*_{T_i,C_i}$ with a simpler parametric model that we denote by f_{T_i,C_i}, where we model T_i as being drawn from some well-known parametric distribution, say Weibull, and we model the conditional pmf of C_i given $K_i = k_i$ as being modeled by our Bernoulli distribution, which has the baked in prior $\Pr\{K_i \in C_i\} = 1$. Ultimately, this provides us with the joint pdf

$$f_{T_i,C_i,K_i}(t_i,c_i,k_i;\boldsymbol{\theta}) = f_{T_i,K_i}(t_i,k_i;\boldsymbol{\theta}) f_{C_i|K_i}(c_i;\boldsymbol{\theta}|k_i).$$

Recall that the Bernoulli model for candidate sets is generally given by

$$\begin{aligned} f_{C_i|K_i}(c_i|k) &= \prod_{j=1}^{m} p_j(k)^{1_{c_i}(j)}(1-p_j(k))^{1-1_{c_i}(j)} \\ &= \left\{\prod_{j\in c_i} p_j(k)\right\}\left\{\prod_{j\notin c_i}(1-p_j(k))\right\}. \end{aligned}$$

where p_j is a function of type $\{1, 2, \ldots, m\} \mapsto [0, 1]$ for $j = 1, \ldots, m$. We are free to choose any such function in this space with the constraint that $p_j(j) = 1$.

If we assume the condition 2.2, then we may impose a constraint like $p_j(k) = \gamma_A$ for $k \in C_i$ and $p_j(k) = \gamma_B$ for $k \notin C_i$, in which case

$$f_{T_i,C_i}(t_i, c_i; \boldsymbol{\theta}) \propto \sum_{k \in c_i} h_k(t; \boldsymbol{\theta}_k) \prod_{j=1}^{m} R_j(t_i; \boldsymbol{\theta}_j),$$

which generates a likelihood function whose maximum is the same as L_R. I think it is an interesting generalization if we do not make this assumption, in which case my derivation of the joint pdf of T_i and C_i represents a potentially new research direction with a lot of interesting possibilities.

Since we do not observe what value K_i realizes in the sample, we find the joint pdf of f_{T_i,C_i} by summing over all the possible values K_i can realize,

$$f_{T_i,C_i}(t_i, c_i; \boldsymbol{\theta}) = \sum_{k=1}^{m} f_{T_i,C_i,K_i}(t_i, c_i, k; \boldsymbol{\theta}).$$

Clearly, we would be much better off if we could observe K_i. In that case, we would just use the joint pdf f_{T_i,K_i} and ignore the candidate set C_i, since the object of statistical interest is the series system, not the candidate sets that are correlated with the series system. But, our data is said to be *masked*, which in this context means we do not know which component failed, we can only narrow it down to some subset we call the candidate set, denoted by C_i.

With this model, we find the parameter values that provide a best-fit for the observed data using the MLE approach. The parameters to estimate are the probability parameters (unless, for some reason, they are given as priors – initially, I had thought this was what you wanted), which we consider to be nuisance parameters, and the object of primary interest, $\boldsymbol{\theta}$.

We can easily make the conditional model C_i given $K_i = k_i$ much more complex in case the simple Bernoulli model is deemed to be incompatible with the data. This gets us into the very interesting topic of model selection, but I don't think we want to go down that path since our time is limited.

That said, let's entertain this idea a bit, since it may help us get a handle on the simpler problem. In what follows, we consider different models of C_i given $K_i = k_i$. The *full* model in this case is given by assigning a potentially different probability to each subset of $\{1, \ldots, m\}$ that includes the (unobserved) failed component label k_i and assigns 0 to every subset that does not include k_i.

For example, if $k_i = 2$ and $m = 3$, then

$$f_{C_i|K_i}(c_i|k = 2) \begin{cases} \gamma_{2|k=2} & \text{if } c_i = \{2\} \\ \gamma_{12|k=2} & \text{if } c_i = \{2, 1\} \\ \gamma_{23|k=2} & \text{if } c_i = \{2, 3\} \\ \gamma_{123|k=2} & \text{if } c_i = \{1, 2, 3\} \\ 0 & \text{otherwise.} \end{cases}$$

There are 2^{m-1} possible candidate sets that include k_i. There are also m values that k_i may take on, and thus there are $m2^{m-1}$ probability parameters to fit in this full model. For small m (for $m = 3$, there are $3 \cdot 2^{3-1} = 12$ parameters), this may not be too burdensome, but the number of parameters grows exponentially as m increases and since the sampling variance of the estimator increases as the number of parameters to estimate increases, we are motivated to seek a reduced model that is sufficiently compatible with the data.

If we have prior information, we can often seek out reduced models that substantially reduce the number of parameters. In the Bernoulli model I described previously,

$$f_{C_i|K_i}(c_i|k) = \prod_{j=1}^{m} p_j(k)^{1_{\{j \in c_i\}}} (1 - p_j(k))^{1_{\{j \notin c_i\}}}$$

where $p_j(k) = 1_{\{j=k\}} + \gamma_j 1_{\{j \neq k\}}$, there are only m probability parameters to estimate, $\gamma_1, \ldots, \gamma_m$, but it may discard a lot of correlations in the data.

The *full* model in our family of Bernoulli candidate models is given by

$$p_j(k) = \begin{cases} \gamma_{j1} & \text{if } k = 1 \\ \gamma_{j2} & \text{if } k = 2 \\ \vdots & \vdots \\ \gamma_{j(j-1)} & \text{if } k = j-1 \\ 1 & \text{if } k = j \\ \gamma_{j(j+1))} & \text{if } k = j+1 \\ \vdots & \vdots \\ \gamma_{j(m-1)} & \text{if } k = m-1 \\ \gamma_{jm} & \text{if } k = m \end{cases}$$

for $j = 1, \ldots, m$.

$p_j(k)$ has $m-1$ probability parameters $(\gamma_{j1}, \ldots, \gamma_{jm})$ for each $j \in \{1, \ldots, m\}$. That means there are $m(m-1)$ probability parameters in the full Bernoulli model. If $m = 3$, that means there are 6 parameters to estimate. In general, for large m, this may also be problematic. Of course, if we have some insight about how the candidate sets are generated, we may be able to drastically reduce the number of parameters, e.g., suppose we have an $m = 3$ electronic series system and components 1 and 2 are on the same circuit board, and a person repaired the system by either replacing the circuit board (meaning that either component 1 or 2 failed, but we do not know which), or replacing component 3 (in which case we know that component 1 caused the failure). In this case, $p_1(1) = p_1(2) = p_2(1) = p_2(2) = 1$, $p_1(3) = p_2(3) = 0$, $p_3(1) = p_3(2) = 0$, and $p_3(3) = 1$. There are, in fact, no parameters to estimate in this case, and furthermore we may use L_R to find the infinite number of points that maximizes it.

So, the Bernoulli model, where we condition C_i on $K_i = k_i$, depends on knowing the joint distribution of K_i and T_i. I'm not sure if you agree with me yet, but this is in fact given by

$$f_{K_i,T_i}(k, t; \theta) = h_j(t; \theta_j) \prod_{k=1}^{m} R_k(t; \theta_k).$$

It is not the conditional pmf of K_i given $T_i = t_i$. That's just

$$f_{K_i|T_i}(k; \boldsymbol{\theta}|t) = \frac{h_k(t; \boldsymbol{\theta}_k)}{\sum_{j=1}^{m} h_j(t; \boldsymbol{\theta}_j)}.$$

If you don't accept my earlier proofs, see [2], which makes the same claim, i.e., $f_{K_i|T_i}(k; \boldsymbol{\theta}|t)$ is another notation for $P_k(t)$ that they use in their paper. I am certainly willing to try to revise my proof to meet your approval, although we may just be satisfied with the following proof.

Proof. In [2], equation (22), $P_k = \frac{h_k(t)}{\sum_{j=1}^{m} h_j(t)}$ is the conditional pmf that component label k is the cause of a series system failure given a system failure at time t. I denote this conditional pmf in my draft paper as $f_{K_i|T_i}(k; \theta|t)$.

The pdf of the series system failure time is given by

$$f_{T_i}(t; \theta) = \left\{ \sum_{j=1}^{m} h_j(t; \theta) \right\} \left\{ \prod_{j=1}^{m} R_j(t; \theta) \right\}.$$

Thus, the joint pdf of K_i and T_i is given by

$$\begin{aligned} f_{K_i,T_i}(k,t;\theta) &= f_{K_i|T_i}(k;\theta|t) f_{T_i}(t;\theta) \\ &= \left\{ h_j(t;\theta_j) \prod_{k=1}^{m} R_k(t;\theta_k) \right\} f_{T_i}(t;\theta) \left\{ \sum_{j=1}^{m} h_j(t;\theta) \right\} \left\{ \prod_{j=1}^{m} R_j(t;\theta) \right\} \\ &= \left\{ h_j(t;\theta_j) \prod_{k=1}^{m} R_k(t;\theta_k) \right\} f_{T_i}(t;\theta) \left\{ \sum_{j=1}^{m} h_j(t;\theta) \right\} \left\{ \prod_{j=1}^{m} R_j(t;\theta) \right\}. \end{aligned}$$

□

Anyway, the closer our fitted parametric model is to the real data generating process, the more accurate inferences will be. Note that our model does not capture many kinds of correlations that may exist in the data generating process for candidate sets. For instance, the candidate sets may be correlated with time. It is relatively straightforward to condition C_i on both $K_i = k_i$ and $T_i = t_i$, e.g., in the Bernoulli model, replace $p_j(k)$ with $p_j(t,k)$, e.g., $p_j(t,k) = 1_{\{j=k\}} + 1_{\{j \neq k\}}\big(1 - \exp(\beta_j t)\big)$.

References

[1] Guess, F. M., Hodgson, T. J. and Usher, J. S. (1991). Estimating system and component reliabilities under partial information on cause of failure. *Journal of Statistical Planning and Inference* **29** 75–85.

[2] Guo, H., Niu, P. and Szidarovszky, F. (2013). Estimating component reliabilities from incomplete system failure data. *Annual Reliability and Maintainability Symposium (RAMS)* 1–6.