Bayesian life-course models revisited

Justin Chumbley[1], Cecilia Potente[1], Wenjia Xu[1], & Michael Shanahan[1]

[1] Jacobs Center for Productive Youth Development, University of Zurich

Author Note

Correspondence concerning this article should be addressed to Justin Chumbley, Jacobs Center for Productive Youth Development, Andreasstrasse 15, CH-8050 Zurich. E-mail: justin.chumbley@pm.me

Abstract

Here we study the relative importance of exposure at different life periods for later health. This question is often framed qualitatively, as a choice between three cases: the *accumulation* (all periods have the same importance), *critical-period* (only one period is important) and *sensitive-period* hypotheses. The latter is a vast composite hypothesis, defined only in opposition to the former two point hypotheses. Building on Madathil, Joseph, Hardy, Rousseau, and Nicolau (2018), we propose two novel Bayesian quantities which a) pit these three broad hypotheses against one another and crucially, in the case of the sensitive hypothesis, b) automatically identify which particular periods are more important.

Keywords: Developmental epidemiology, regression, life course models, ...

Word count: X

Bayesian life-course models revisited

Potential reviewers

- Madathil, ...

Introduction

We assume that a subject's exposure history $\mathbf{x} = (x_1, ..., x_T)$ over T periods impacts their subsequent outcome Y without time-dependent confounding according to a generalized linear model with conditional expectation of the form $E(Y|\mathbf{x}) = g(\mathbf{x}\boldsymbol{\theta}')$[1]. The parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_T)$ captures the impact of exposure x at each time period $t = 1, ..., T$ on outcome Y, and we would like to contrast these effects to one another to see which period(s), if any, matter more. Madathil et al. (2018) pursue an interesting non-linear reparameterization $\boldsymbol{\theta} = \delta\mathbf{w}$ where $\delta \in \mathbb{R}$ and $\mathbf{w} \in \Delta^T$. Here $\Delta^T := \{(w_1, .., w_T) \in \mathbb{R}^T : w_t > 0, \sum_t w_t = 1\}$ is the T-part regular simplex, which has $T - 1$ degrees of freedom. This parameterization forces all non-zero weights to share the same sign, which captures the typical epidemiological setting, and gives the parameters nice interpretations: w_t now encodes the relative effect of time period t while δ is "the total lifetime effect"[2].

Madathil et al. (2018) assume a uniform prior on the weights $\mathbf{w}$. To compare the three broad hypotheses above, they ask whether their multivariate posterior 95% credible regions exclude or include (cover) the accumulation or critical hypotheses (points): these points are respectively a) $w_t = 1/T$ for all t, for example $\mathbf{w} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ in the $T = 3$ period situation, and b) $w_t = 1$ for some unique t, for example $\mathbf{w} = (0, 0, 1)$. Their strategy accommodates a wide variety of exposure and outcome variables, missing values and measurement errors and offers other welcome advantages of the Bayesian approach. Yet some obvious issues remain

[1] Here g is some link function and linear covariates have been omitted without loss of generality.
[2] Although any non-identity link function g will also influence the interpretation of $\mathbf{w}$.

due to the limitations of continuous, multidimensional credible sets, which are particularly acute on bounded parameter spaces like Δ^T. Depending on the care taken in their construction and interpretation, multivariate credible sets may falsely exclude the critical model, which is found on the boundary of parameter space. They may also falsely include the accumulation model, if practitioners erroneously conclude that marginal covering implies joint covering[3]. It is evident that some univariate alternative to multivariate credible sets would be preferable. More importantly, even when the accumulation and critical point hypotheses are appropriately excluded by a posterior credible set, no conventional credible set can answer the most basic question: "*which specific periods are more sensitive than which?*".

In the spirit of Madathil et al. (2018)'s "continuous model expansion", we address these issues by constructing credible sets on two *transformations* of their posterior distribution $p(\mathbf{w}|y)$. The first seeks a simple univariate means to compare the three hypotheses, the second to decompose and interpret the vast sensitive-hypothesis. The first transformation calculates the greatest difference between any two period's importance, i.e. the range of component parameters $\mathbf{w} = (w_1, ..., w_T)$. This range characterizes the three hypotheses: it equals zero and one for the accumulation and critical hypotheses respectively, being strictly between zero and one for any sensitive model. It therefore contains all the information required to choose between hypotheses and offers a practical alternative to conventional model selection. Alternatively, it may simply just be viewed as a quantitative index of how similar the weights are across time periods. The second transformation ranks periods by importance, assessing which periods are more "sensitive" (have larger weight) than which others. This permits a practitioner to give a numeric probability to the conclusion that say "the first two periods of life matter more than any subsequent period".

[3]A multivariate set S may exclude a point p even while all of P's lower dimensional projections include all of p's projections.

Rationale

Our first transformation is $\phi : \Delta^T \to \Delta^2$ from the T-part simplex to the 2-part simplex or unit interval $[0, 1]$, defined by $\phi : \mathbf{w} \mapsto (\vee\mathbf{w} - \wedge\mathbf{w})$. Here $\vee$ is the max operation that extracts the magnitude of the largest component of $\mathbf{w}$, dually for the min operation $\wedge$. The function ϕ therefore gives the range of components of $\mathbf{w}$: the Euclidean distance between the maximum and minimum values of the components of $\mathbf{w}$. Note that $\phi((\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) = \frac{1}{3} - \frac{1}{3} = 0$ for the accumulation hypothesis and $\phi((1, 0, 0)) = \phi((0, 1, 0)) = \phi((0, 0, 1)) = 1 - 0 = 1$ for any critical hypothesis. Note also for example that $\phi(0.2, 0.7, 0.1)) = 0.6 = 0.7 - 0.1$. This particular ϕ value of 0.6 represents a sensitive hypothesis because it is neither 0 nor 1 (pure accumulation or critical hypotheses). Other solutions to $\phi(\mathbf{w}) = 0.6$ represent sensitive hypotheses which are equivalent to $(0.2, 0.7, 0.1)$ in as much as they have the same range, including for example $\phi((0.8, 0.2, 0)) = 0.6$.

More generally, any choice of two distinct thresholds $a, b \in [0, 1]$ partition the $\mathit{codomain}(\phi) = [0, 1]$ into three sets practically equivalent to the $\mathit{accumulation} = [0, a]$, $\mathit{sensitive} = (a, b)$ and $\mathit{critical} = [b, 1]$ hypotheses respectively[4]. The strictest definition of these three hypotheses arises in the limit that a tends to 0 and b tends to 1. Note that $p(\mathit{sensitive\ or\ accumulation\ or\ critical}|y) = 1$ so this approach effectively just discretizes $p(\phi|y)$ or equivalently[5] $p(\mathbf{w}|y)$. If $p(\mathit{sensitive}|y) \geq 0.95\%$, for example, we may conclude that the sensitive model is credible in practice.

Our second transformation then seeks to whittle down this broad sensitive period

[4]In principle this can be further generalized to include any finite number of break points $a_i \in [0, 1]$, yielding intermediary classes.

[5]Note that such a partition of the codomain exactly corresponds to a partition of $\mathit{domain}(\phi) = \Delta^T = M_0 \cup M_1 \cup M_2 := \{\mathbf{w} : 0 \leq \phi \leq a\} \cup \{\mathbf{w} : a < \phi < b\} \cup \{\mathbf{w} : b \leq \phi \leq 1\}$. Just as the posterior density of $p(\phi|y)$ at the value 0.6 reflects the total density over all these solutions $\{\mathbf{w} : \phi(\mathbf{w}) = 0.6\}$, our tripartition of $\Delta^T = M_0 \cup M_1 \cup M_2$ effectively discretizes continuous posterior density $p(\mathbf{w}|y)$ satisfying $1 = \int_{\Delta^T} p(\mathbf{w}|y)$ into discrete density $P(M_i|y)$ satisfying $1 = \sum_{i=1}^{3} P(M_i|y)$.

hypothesis into something more specific. This transformation is $f : \Delta^T \to \mathcal{S}_T$ from the simplex to the set of all full rankings - i.e. permutations - of the labels {"w_1","w_2",...,"w_T"} or equivalently of $\{1, 2, ..., T\}$. In particular, f assigns each vector $\mathbf{w}$ the full ranking of it's components. For example, it maps the point $\mathbf{w} = (0.2, 0.7, 0.1)$ to the full ranking $w_3 < w_1 < w_2$: we will henceforth use the terser notation $3|1|2$ for such full rankings. Other solutions to $f(\mathbf{w}) = 3|1|2$, such as $\mathbf{w} = (0.3, 0.5, 0.2)$, represent sensitive hypotheses which are equivalent to $(0.2, 0.7, 0.1)$ in as much as the relative importance of periods is exactly the same. In this way we can again discretize continuous posterior density $p(\mathbf{w}|y)$, this time into the discrete density $P(\pi|y)$ over $T!$ full rankings π such as $3|1|2$ satisfying $1 = \sum_{\pi \in \mathcal{S}_T} P(\pi|y)$. This enables us to answer directly whether the most probable ranking of periods by importance is say $\pi = 3|1|2$, or whether say $p(\pi|y) \geq 0.95\%$. In this way we gain insight into our multivariate posterior $p(\mathbf{w}|y)$ without the inconvenience of complicated continuous multivariate credible sets.

The function f also permits us to define more general, *partial rankings* such as $3, 1|2$ and calculate their posterior probability. Such partial rankings denote collections of full rankings consistent with a weaker conclusion than a single full ranking on it's own. For example, if $\pi = 1, 3|2$ and $p(\pi|y) \geq 0.95\%$ then we can say with 95% posterior credibility that period 2 is more important than the other two periods, even though we can say nothing about the relative importance of these latter. To elaborate, the partial ranking $1, 3|2$ represents points $\mathbf{w}$ that can be ranked *either* as $w_3 < w_1 < w_2$ as previously, or as $w_1 < w_3 < w_2$: it therefore encodes points $\mathbf{w}$ for which w_2 is unambiguously the most important or largest period. However this ordering is *partial* because either $w_3 < w_1$ or $w_3 < w_1$. The set of all symbols like $1, 3|2$ which include T integers separated by either a bar "|" or a comma "," give a space of statements that are both readily interpretable and can be assigned posterior probability. In this framework, the β% finest credible rank, which we denote $\mathcal{C}_\beta$, is the smallest such set with β% posterior probability.

Goals of simulation

We asked whether the univariate credible interval for $\delta|y$ appropriately excluded zero, i.e. correctly inferred whether *any* time period is relevant for the outcome. If the answer to this is positive, it makes sense to broadly examine the three competing hypotheses of the introduction via the posterior distribution $\phi|y$. If $\phi|y$ additionally supports a sensitive hypothesis, it makes sense to break this down by examining $f|y$ and the probable ordering of weights.

The objectives of the simulation study were therefore: (a) to assess whether $\delta|y$ appropriately excluded zero; (b) to assess whether the range $\phi|y$ successfully discriminates between the accumulation, critical and sensitive hypotheses; (c) to assess whether rank $f|y$ succeeds in identifying the correct ranking of time periods by their sensitivity.

Aim (c) is more elaborate, so we seek to answer it in two stages as described next. Knowing the simulated ground truth $\mathbf{w}^*$ and its corresponding true full ranking $f(\mathbf{w}^*)$ our principle questions concern it's relation to the inferred *finest β% credible ranking* which we denote $\mathcal{C}_\beta$.

1) Is $\mathcal{C}_\beta$ consistent with the true full ranking, i.e. $f(\mathbf{w}^*) \in \mathcal{C}_\beta$? We say that $\mathcal{C}_\beta$ is inconsistent if, for example, it asserts $w_2 < w_4$ while the underlying truth is $w_4 < w_2$. Otherwise it is consistent.

2) how much "information" does $\mathcal{C}_\beta$ retain? Here we use $q = r/r^*$ with values between 0 to 1 to measure the quality of $\mathcal{C}_\beta$, where r is the number of distinctions - inequalities or bars "|" - in $\mathcal{C}_\beta$ and r^* the true number in $f(\mathbf{w}^*)$. Larger q therefore means a more informative inference.

The first question expresses the minimal requirement that $\mathcal{C}_\beta$ does not contradict the truth. But we additionally want $\mathcal{C}_\beta$ to be as informative as possible, ideally faithfully retaining *all* distinctions made in the true ranking $f(\mathbf{w}^*)$. Hence question two.

Simulation parameters

Our simulation fully reproduced and extended that of Madathil et al. (2018). Namely, we simulated a three-period life course study assuming no measurement error in the variables. In particular, for the i'th participant we sampled three Gaussian exposure variables $\mathbf{x}_i = (x_1, x_2, x_3)$ with a correlation of 0.7 and 0.49 between adjacent and non-adjacent measures, respectively. Datasets were simulated for all combinations of the four life course hypotheses and three sample sizes ($n = 700, 1500, 3000$). The ground truth weight values of the simulation, denoted with an asterix "$*$", were: (i) pure accumulation hypothesis $\mathbf{w}_i^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$; (ii) monotonic sensitive period model with weights $\mathbf{w}_{ii}^* = \frac{1}{1+2+3}(1, 2, 3)$; (iii) first life period as a sensitive period $\mathbf{w}_{iii}^* = (0.75, 0.2, 0.05)$; (iv) third life period as a critical period $\mathbf{w}_{iv}^* = (0, 0, 1)$.

We extended Madathil et al. (2018)'s 3 period simulation to 5 and 7 periods in the natural way as follows. The accumulation model (i) above was generally $\mathbf{w}_i^* = (\frac{1}{T}, ..., \frac{1}{T})$; (ii) was generalized in the obvious way to $\mathbf{w}_{ii}^* = \frac{1}{\sum_{i=1}^{T} t}(1, 2, ..., T)$ and (iii) and (iv) were padded with zeros, being for example $\mathbf{w}_{iii}^* = (0.75, 0.2, 0.05, 0, 0, 0, 0)$ and $\mathbf{w}_{iv}^* = (0, 0, 0, 0, 0, 0, 1)$ respectively in 7 dimensions.

We independently varied the lifetime effect δ^* between 0,1 and 2: these fixed underlying settings (estimands) are again denoted "$*$", to distinguish them from the posterior inferred counterparts. Given $\delta^*, \mathbf{w}^*$, we then generated $y_i = \delta^* \sum_{j=1}^{T} x_{ij} w_j^* + \epsilon_i$ with independent $\epsilon_i \sim N(0, 1)$, for $i = 1, ..., n$.

Prior, likelihood and posterior

In accordance with the data generating model above, we used Bayesian linear regression for inference, as discussed below. Analogous analyses easily pertain for other variants of the generalized linear model. We followed Madathil et al. (2018) in their choice of a uniform prior over all Δ^T, namely a non-informative Dirichlet prior for weights $p(\mathbf{w}) = Dirichlet(\mathbf{w}|\mathbf{1})$, where $\mathbf{1}$ is a vector of T ones, and a weakly informative Cauchy prior on the lifetime effect $p(\delta) = Cauchy(\delta|0, 2.5)$. In cases where there is plausible justification for bias towards the accumulation or critical models, the hyperparameter can be generalized to $c\mathbf{1}$, with $c > 0$. Then it is well-known from the properties of the Dirichlet that choosing $c < 1$ implies a bias toward the critical hypothesis, and $c > 1$ a bias towards accumulation.

We then used the No-U-Turn MCMC sampler as implemented in STAN (Carpenter et al., 2017) to acquire 50k marginal posterior samples of $\delta|y$ and $\mathbf{w}|y$. Having performed standard convergence tests, we examined $\delta|y$, and derived $\phi|y$ and $f|y$ by applying ϕ, f to each point in our posterior sample $\mathbf{w}|y$.

Results of simulation

Posterior lifetime effect δ and range ϕ

In every simulation the univariate credible interval for $\delta|y$ appropriately excluded zero. Thus $\delta|y$ is a faithful omnibus measure. In simulations with evidence of non-zero $\delta|y$, we proceeded to examine the posterior range $\phi|y$. The violin plots in Figure 1 depict this posterior distribution of the range $\phi|y$ under each setting of $\mathbf{w}^*$ the ground truth, the total life-time effect δ^*, the number of periods T, and the sample size n. The x-axis of each plot distinguishes the 4 settings i-iv for $\mathbf{w}^*$ defined in section “Simulation parameters”.

Note that the posterior distribution $\phi|y$ approaches the truth $\phi(\mathbf{w}^*)$ with increasing δ^* and n. To convert these plots into a more formal comparison between the *accumulation*, *critical-period* and *sensitive* hypotheses, we choose $a = 0.15$ and $b = 1 - a$ to define regions practically equivalent to each of these three cases, as discussed in section "Rationale". Table 1 gives the ensuing confusion matrix which relates the inferred hypothesis (column variable) to the underlying truth (row variable). Recall that i refers to simulations where the accumulation model is true, ii, iii indicate the sensitive models and iv indicates the critical model. The table illustrates that we could always faithfully recover the ground truth, albeit at the price of occasionally confessing ignorance.

Table 1

Confusion matrix for a choice between the accumulation (a), sensitive (s) and critical (c) hypotheses, among simulations with evidence of a non-zero lifetime effect. We made a conclusive choice between these cases whenever one of them had probability greater than 0.9 posterior probability, otherwise our inference was considered inconclusive/unknown (u).

	a	s	c	u
i	14	0	0	4
ii	0	18	0	0
iii	0	17	0	1
iv	0	0	14	4

When the sensitive hypothesis was credible, i.e. the 18+17 cases in column s of Table 1, it makes sense to ask which periods are more or less sensitive. For these cases we therefore calculated the finest partial ranking of parameters, as discussed next.

Table 2

The finest credible ranking from a representative subset of simulations. Columns record the true model (ii or iii defined above), sample size, number of periods, true order and infered ranking fcr, whether the infered ranking (fcr) violates or contradicts the true order, and the fraction of distinctions q that fcr preserves.

true_model	n_samples	n_periods	truth	fcr	violate	q
ii	3000	5	1|2|3|4|5	1|2|3|4|5	FALSE	1.000
iii	3000	5	4,5|3|2|1	4,5|3|2|1	FALSE	1.000
iii	700	7	4,5,6,7|3|2|1	7,6,4,5,3|2|1	FALSE	0.667
ii	3000	5	1|2|3|4|5	1|2|3|4|5	FALSE	1.000
iii	1500	5	4,5|3|2|1	4,5,3|2|1	FALSE	0.667
iii	1500	7	4,5,6,7|3|2|1	7,6,5,4,3|2|1	FALSE	0.667
iii	1500	7	4,5,6,7|3|2|1	5,6,4,7|3|2|1	FALSE	1.000
iii	700	7	4,5,6,7|3|2|1	6,5,4,3,7,2|1	FALSE	0.333
iii	1500	3	3|2|1	3|2|1	FALSE	0.667
ii	1500	3	1|2|3	1|2|3	FALSE	1.000

Posterior finest credible rank $\mathcal{C}_\beta$

Figure 2 illustrates our recursive scheme for whittling down the sensitive hypothesis. In particular, it gives the "cumulative density function" of a special nested increasing set of subsets of Δ^T which leads to what we call $\mathcal{C}_{90\%}$, the finest 90% credible ranking. The candidate partial ranking at each step in the sequence from left to right is the most credible (maximum probability) coarsening of the preceding candidate. In practice, the posterior probability of each candidate ranking was estimated as the fraction of posterior samples satisfying the relevant inequalities. We selected $\mathcal{C}_{90\%}$ as the first candidate that exceeds the desired credibility threshold of 0.9, depicted in Figure 2 as a black horizontal line. The column "fcr" of Table 2 gives some examples of this inferred finest credible ranking $\mathcal{C}_{90\%}$ across different simulations. This can be compared with the "truth" column to answer the two specific questions posed in section "Goals of Simulation":

1) We found that our inferred partial rank $\mathcal{C}_\beta$ rarely violated the ground truth. Such a violation occurred in 0 percent of the simulations.

2) On average over all simulations 0.71 % of the distinctions were preserved. Table 3 shows that q, the proportion of distinctions preserved in $\mathcal{C}_\beta$, increased with the simulated sample size.

Table 3

The mean proportion q of distinctions preserved by the posterior credible ranking increased with the simulated sample size.

n	q
700	0.52
1500	0.72
3000	0.89

Discussion

We have offered greater insight into the relative contribution of different life periods for some subsequent outcome by considering judicious reparameterizations of the simple model described in Madathil et al. (2018). Our analysis inherits all the desirable features of their Bayesian approach, but offers new advantages. We ease intuition and reportability by offering the scientist a way to draw principled *qualitative* conclusions, both on the best "model" - accumulation, critical and sensitive - and in the latter case, which periods are relatively more sensitive. This is achieved by relaxing the dependence on both continuous multivariate confidence sets and the chosen metric on parameter space (chosen to be Euclidean metric as opposed to say Hilberts projection metric or the Aitchison metric).

The ease with which we can flexibly pose and answer such general questions is a key motivation for taking the Bayesian path. By comparison, the frequentist linear modelling framework (Rosenthal, Rosnow, & others, 1985; Rosenthal, Rosnow, & Rubin, 2000) is limited to rejecting sensitivity hypotheses which can be cast as linear contrasts $\mathbf{w}^T\mathbf{c}$ for some *fully pre-specified* contrast vector $\mathbf{c}$. This traditional workhorse can therefore reject, for example, a fully specified linear or exponential sensitivity hypothesis[6]. Yet the science rarely justifies committing to a linear or exponential trend, let alone a strong hypothetical specification of α. Our proposed approach, which is not limited to rejecting null hypotheses, can positively accept say the much weaker claim that sensitivity decreases in some monotone fashion $w_1 > w_2 > ... > w_T$. This includes the linear and exponential special cases above but is less restrictive (and therefore more credible) because it doesn't insist on a particular functional form for the change in sensitivity over time. More generally still, because sensitivity may vary arbitrarily the appropriate conclusion in a given application may actually transpire to be say $w_2 < \{w_3, w_1\} < w_4$ or $\{w_3, w_1\} < \{w_1, w_4\}$, etc. This lead us to

[6]Such as: sensitivity linearly decays as $w_{t+1} = w_t + \alpha$ with fixed, hypothetical $\alpha < 0$, or exponentially as $w_t = \alpha^t$ with $0 < \alpha < 1$. In the former, $w_{t+1} - w_t = \alpha$ and in the latter $w_{t+1}/w_t = \alpha$ for all $t < T$

develop the notion of the finest credible ranking $\mathcal{C}$ which may be derived automatically from the data.

Prior on ϕ and f. Conventional model selection would explicitly mix point mass priors on the critical and accumulation points with a continuous prior over the sensitive hypothesis. It is partly to avoid artificially distinguishing the two point hypotheses in this way - and necessarily increasing their plausibility - that we follow Madathil et al. (2018) in their choice of a uniform prior over all Δ^T. Note however that prior uniformity of $\mathbf{w}$ does not imply prior uniformity over the critical, accumulation and sensitive hypotheses. Rather, it implies a natural "bias" in favor of the sensitive hypothesis. This is because the critical or accumulation hypotheses occupy less of Δ^T, and therefore are assigned proportionally less prior probability. Technically, a strict interpretation of the critical or accumulation hypotheses as *points* implies that they must receive zero prior (and posterior) probability under any continuous distribution: only volumes have proper probability, and the sensitive hypothesis naturally comprises the entire volume of simplex Δ^T. Analogously, prior uniformity of $\mathbf{w}$ obviously does not generally imply uniformity of the range ϕ of components of $\mathbf{w}$. This reflects a more general issue in Bayesian statistics, often discussed in connection with Jeffries priors, e.g. (Gelman et al., 2013). Our simulations have shown that these subtleties can be overlooked in practice: with enough data the posteriors support faithful reverse inferences.

Conversely, note that each constituent full ranking of the sensitive hypothesis, say $3|1|2$ or $3|2|1$ occupies equal volume of Δ^T. Uniformity of $p(\mathbf{w})$ is therefore preserved by the rank transformation f. In particular, each full ranking is assigned $1/T!$ prior probability. A general partial ranking is assigned prior probability $k/T!$, where k is the number of underlying full rankings that comprise the partial ranking. For example, partial ranking $2|1,3$ has $k = 2$ and prior equaling $2/3!$ because $2|1,3 := 2|1|3$ *or* $2|3|1$.

Future work. Future work should extend our ideas to time-dependent parameters of more explicitly causal models, e.g. (Robins, Hernan, & Brumback, 2000; VanderWeele, Hernán, Tchetgen Tchetgen, & Robins, 2016)....

Acknowledgements

We would like to thank Margaret Bellamy.

REFERENCES

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman; Hall/CRC.

Madathil, S., Joseph, L., Hardy, R., Rousseau, M.-C., & Nicolau, B. (2018). A bayesian approach to investigate life course hypotheses involving continuous exposures. *International Journal of Epidemiology*, *47*(5), 1623–1635.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. LWW.

Rosenthal, R., Rosnow, R. L., & others. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. CUP Archive.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press.

VanderWeele, T. J., Hernán, M. A., Tchetgen Tchetgen, E. J., & Robins, J. M. (2016). Re: Causality and causal inference in epidemiology: The need for a pluralistic approach. *International Journal of Epidemiology*, *45*(6), 2199–2200.

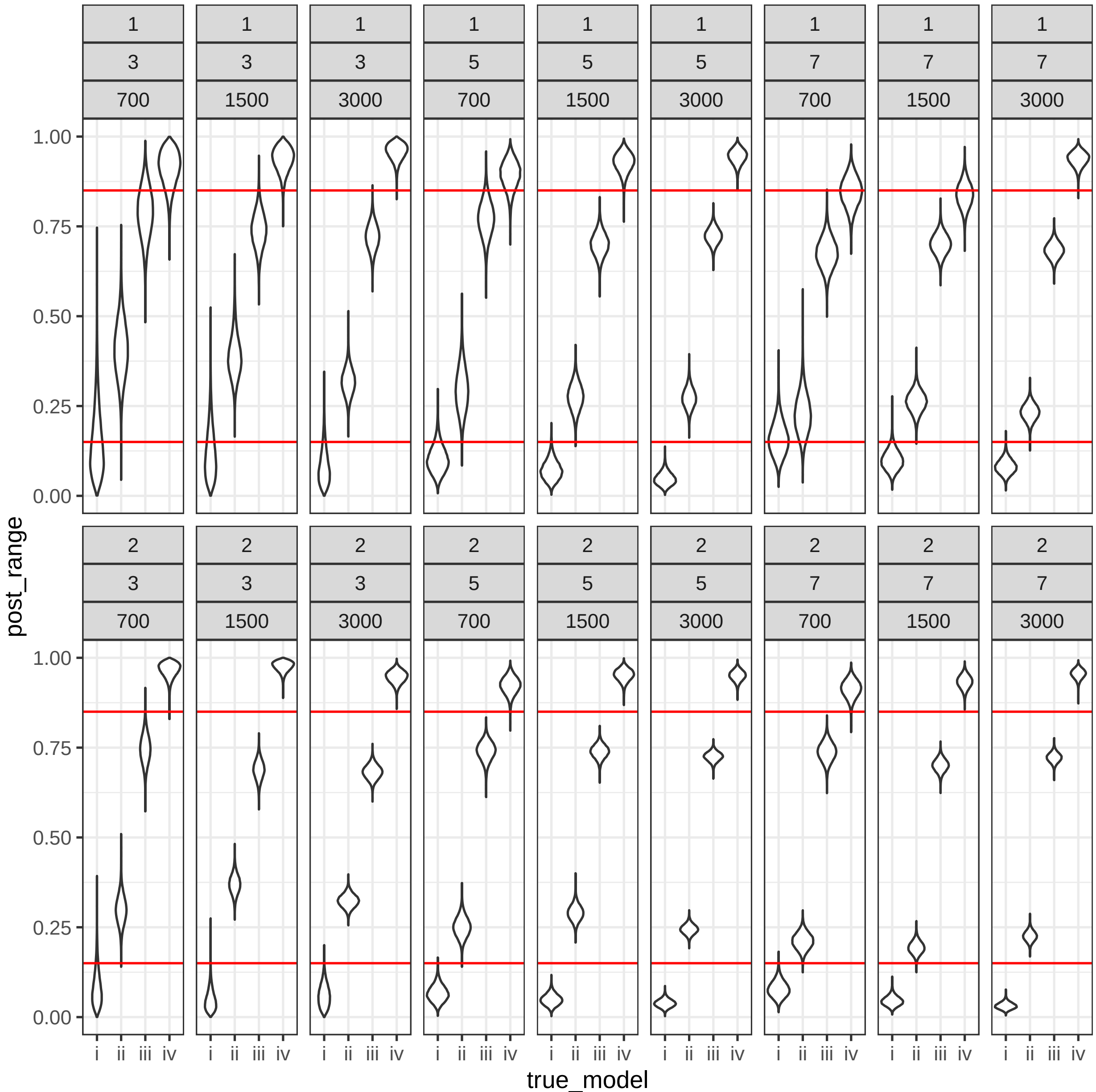

Figure 1. Posterior distribution of the range $\phi|y$ under different simulation conditions. Each cell is labeled (from top to bottom) with an integer giving the ground truth of the life-time effect δ^*, the number of periods T, and sample size n.

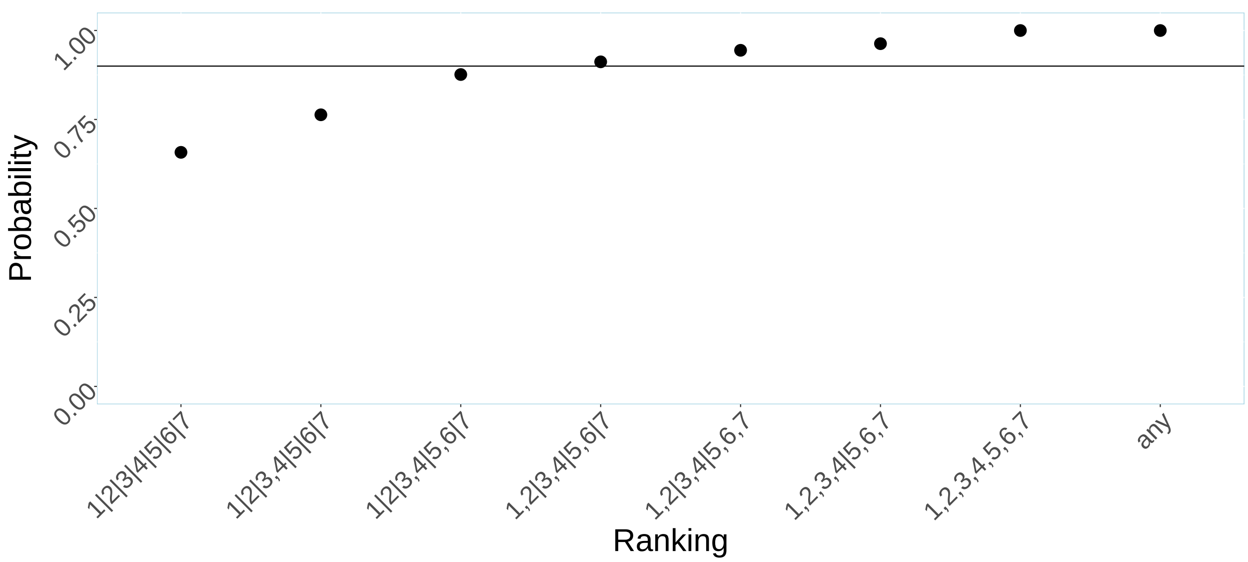

Figure 2. Posterior cumulative density over increasingly coarse partial rankings. The ground truth in this simulation was 1|2|3|4|5|6|7, an instance of model *ii* with monotonically increasing sensitivities. Progressing from left to right across the x axis rankings become coarser by the loss of one distinction ("|"). The choice of which distinction is weakest is determined by a recursive maximization scheme. The finest (leftmost) ranking satisfying criterion, e.g. 90% credibility, may be reported.