Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Supplementary Material for Efficient Computation of High-Dimensional Penalized Piecewise Constant Hazard Random Effects Models

Hillary M. Heiling[1] | **Naim U. Rashid**[1,2] | **Quefeng Li**[1] | **Xianlu L. Peng**[2] | **Jen Jen Yeh**[2,3,4] | **Joseph G. Ibrahim**[1]

[1] Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina, United States of America

[2] Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, North Carolina, United States of America

[3] Department of Surgery, University of North Carolina at Chapel Hill, North Carolina, United States of America

[4] Department of Pharmacology, University of North Carolina at Chapel Hill, North Carolina, United States of America

Correspondence

Corresponding author Hillary Heiling.
Email: hillary_heiling@dfci.harvard.edu

Abstract

Identifying and characterizing relationships between treatments, exposures, or other covariates and time-to-event outcomes has great significance in a wide range of biomedical settings. In research areas such as multi-center clinical trials, recurrent events, and genetic studies, proportional hazard mixed effects models (PHMMs) are used to account for correlations observed in clusters within the data. In high dimensions, proper specification of the fixed and random effects within PHMMs is difficult and computationally complex. In this paper, we approximate the proportional hazards mixed effects model with a piecewise constant hazard mixed effects survival model. We estimate the model parameters using a modified Monte Carlo Expectation Conditional Minimization algorithm, allowing us to perform variable selection on both the fixed and random effects simultaneously. We also incorporate a factor model decomposition of the random effects in order to more easily scale the variable selection method to larger dimensions. We demonstrate the utility of our method using simulations, and we apply our method to a multi-study pancreatic ductal adenocarcinoma gene expression dataset to select features important for survival.

KEYWORDS

factor model, mixed effects, piecewise constant hazard, survival analysis, variable selection

1 | SUPPLEMENTARY MATERIAL: ADDITIONAL PHMMPEN_FA SIMULATION AND ALGORITHM DETAILS

1.1 | B matrices used in simulations

Equations (1) and (2) give the transpose of the first 6 rows of the deterministic 'small' and 'moderate' $\boldsymbol{B}$ matrices used in the piecewise constant hazard simulations in Section 3 of the main manuscript. All other $p - 5$ rows of the $\boldsymbol{B}$ matrices were set to 0, where $p \in \{100, 500\}$ in the piecewise constant hazard mixed effects simulations.

Equation (3) gives the transpose of the first 6 rows of the deterministic 'moderate' $\boldsymbol{B}$ matrix used in the piecewise constant hazard simulations in Section 3.4 of this Supplementary Material. All other $p-5$ rows of the $\boldsymbol{B}$ matrix (with $p = 100$) were set to 0.

$$\boldsymbol{B}_{moderate}^T = 0.75 \times \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 \end{bmatrix} \tag{1}$$

$$\boldsymbol{B}_{small}^T = 0.50 \times \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 \end{bmatrix} \tag{2}$$

Abbreviations: GLMM, generalized linear mixed model; MCECM, Monte Carlo expectation conditional minimization; PHMM, proportional hazards mixed model

$$\boldsymbol{B}^T_{r=5,moderate} = 0.65 \times \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & -1 & 2 & 0 \\ 1 & 0 & 1 & -1 & 0-2 & \end{bmatrix} \qquad (3)$$

1.2 | Model selection

This section provides details on how the **phmmPen_FA** algorithm selects the optimal tuning parameter combination. In all simulations and analyses discussed in this paper, we set the random effects equal to the fixed effects (i.e. $q = p + 1$, where the additional 1 accounts for a random intercept) and let the algorithm select the fixed and random effects.

The algorithm runs a computationally efficient two-stage approach to pick the optimal set of tuning parameters. In the first stage of this approach, the algorithm fits a sequence of models where the fixed effect penalty is kept constant at the minimum value of the fixed effects penalty sequence, labeled here as $\lambda_{0,min}$, and the random effects penalty proceeds from the minimum random effect penalty, labeled $\lambda_{1,min}$, to the maximum value $\lambda_{1,max}$. The best model from this first stage is then identified using the BIC-ICQ criterion.[1] This first stage identifies the optimal random effect penalty value, $\lambda_{1,opt}$. In the second stage, the algorithm fits a sequence of models where the random effects penalty is kept fixed at $\lambda_{1,opt}$ and the fixed effects penalty proceeds from its minimum value $\lambda_{0,min}$ to its maximum value $\lambda_{0,max}$. The overall best model is chosen from the models in the second stage.

We have found this two-stage model selection approach to work very well in practice (see Section 3 of the main manuscript and Section 3 of this Supplementary Material for performance results).

1.3 | Tuning parameter selection

The default maximum penalty, labeled here as λ_{max}, was calculated as the penalty that would penalize all of the fixed effects to 0 when no random effects are in the model. We used code from the **ncvreg** R package[2] to calculate this value.

For all piecewise constant hazard mixed effect variable selection simulations and case study analyses where the total number of predictors was 100, we used the following sequence of penalties for both the fixed effects and the rows of the $\boldsymbol{B}$ matrix: a sequence of 10 penalties from $0.05\lambda_{max}$ to λ_{max}, with penalty values equidistant from each other on the log scale.

For all piecewise constant hazard mixed effects variable selection simulations where the total number of predictors was 500, we used the following sequence of penalties: a sequence of 10 penalties from $0.25\lambda_{max}$ to λ_{max} for the fixed effects, and a sequence of 10 penalties from $0.10\lambda_{max}$ to λ_{max} for the rows of the $\boldsymbol{B}$ matrix, with penalty values equidistant from each other on the log scale.

For the case study that performed variable selection on a piecewise constant hazard mixed effects model with $p = 168$ TSP covariates, we used a sequence of 10 penalties from $0.10\lambda_{max}$ to λ_{max} for both the fixed effects and the rows of the $\boldsymbol{B}$ matrix, with penalty values equidistant from each other on the log scale.

1.4 | Initialization and convergence for phmmPen_FA

The fixed effects $\tilde{\boldsymbol{\psi}}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$ and random effects covariance terms $\boldsymbol{b}^{(0)}$ are initialized at the MCECM iteration $s = 0$ in one of two ways. We discuss first the initialization procedure used when the method **phmmPen_FA** is used to fit the first model in the sequence of models fit for variable selection. In this scenario, the fixed effects $\tilde{\boldsymbol{\psi}}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$ are initialized by fitting a 'naive' piecewise constant hazard model using the coordinate descent techniques of Breheny and Huang[2] assuming no random effects.

Based on the initialized fixed effects $\boldsymbol{\beta}^{(0)}$, the predictors with non-zero initialized fixed effects are also initialized to have non-zero random effects (i.e. the corresponding rows of the $\boldsymbol{B}$ matrix are set to non-zero values), and predictors with zero-valued initialized fixed effects are initialized to have zero-valued random effects (i.e. the corresponding rows of the $\boldsymbol{B}$ matrix are set to zero). By default, the starting $\boldsymbol{B}$ matrix elements are initialized as $\sqrt{0.02/r}$, where r is the estimated number of latent factors. The corresponding initialized covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T$ will have all non-zero elements equal to 0.02. We found that increasing these initial covariance element values tended to increase the false positives of the fixed and random effects of the algorithm.

The E-step MCMC chain of the sample of the posterior density $\phi(\boldsymbol{\alpha}_k|\boldsymbol{\omega}_{k,o};\boldsymbol{\theta}^{(s)})$ for groups $k = 1, ..., K$ is initialized in iteration $s = 1$ with random draws from the standard normal distribution. For all following iterations $s > 1$, the MCMC chain is initialized with the last draw from the previous iteration $s - 1$. At iteration $s = 0$, we sample $M = 100$ posterior samples from each group, and M increases to a max of 500 as the iteration number s increases.

When the algorithm fits consecutive models in the sequence of models fit for variable selection (i.e. the second model or the third model, and so on), we initialize the coefficients of models with the coefficients estimated in the previous model results. After the first model is fit in the variable selection procedure, the fixed effects, random effects covariance matrix, and random effects MCMC chain are initialized using results from a previous model fit. We found that progressing through penalty values from the minimum penalty to the maximum penalty (as discussed above and in Sections 1.2 and 1.3) significantly improved initialization of subsequent models, as opposed to proceeding from the maximum penalty to the minimum penalty as some other fixed effects only penalization methods do.[3,2]

The EM algorithm is considered to have converged when the following condition is met at least 2 consecutive times (default) or until the maximum number of EM iterations (25) is reached:

$$||(\tilde{\boldsymbol{\psi}}^{(s)T}, \boldsymbol{\beta}^{(s)T}, \boldsymbol{b}^{(s)T})^T - (\tilde{\boldsymbol{\psi}}^{(s-t)T}, \boldsymbol{\beta}^{(s-t)T}, \boldsymbol{b}^{(s-t)T})^T||_2^2 / d_n^{s-t} < \epsilon_{EM} \tag{4}$$

where the superscript $(s - t)$ indicates t EM iterations back (default $t = 2$), $||.||_2^2$ represents the L_2 norm, and d_n^{s-t} equals the total number of non-zero $(\tilde{\boldsymbol{\psi}}^T, \boldsymbol{\beta}^T, \boldsymbol{b}^T)^T$ coefficients in iteration $(s - t)$. In other words, the algorithm computes the average Euclidean distance between the current coefficient vector $(\tilde{\boldsymbol{\psi}}^T, \boldsymbol{\beta}^T, \boldsymbol{b}^T)^T$ and the coefficient vector from t EM iterations back and compares it with $\epsilon_{EM} = 0.0015$.

The M-step algorithm is considered to have converged when the following condition is met or until the maximum number of iterations (50) is reached:

$$\max\{\max_j|\tilde{\psi}_j^{(s,g+1)} - \tilde{\psi}_j^{(s,g)}|, \max_l|\beta_l^{(s,g+1)} - \beta_l^{(s,g)}|, \max_{t,h}|b_{th}^{(s,g+1)} - b_{th}^{(s,g)}|\} < \epsilon_m, \tag{5}$$

where b_{th} is an individual element of $\boldsymbol{b}_t$, which is the t-th row of the $\boldsymbol{B}$, and g represents the iteration number of the M-step. The value of ϵ_m was set to 0.001.

2 | SUPPLEMENTARY MATERIAL: ADDITIONAL CASE STUDY INFORMATION

2.1 | Case study: Study information and data processing

In this section, we provide more information about the individual studies contained within the dataset and describe how we set up the data for the case study analyses in Section 4 of the main paper. More complete coding details are provided in the GitHub repository https://github.com/hheiling/paper_phmmPen_FA.

The studies used in these analyses are summarized in Supplementary Material Table 1, which contains the gene expression platform information (RNA-seq vs microarray), the gene set sizes, the sample sizes, the number of events, and the corresponding citations for each of the studies.

Dataset	Platform	Gene Set Size	Sample Size	# Events	Citation
Aguirre	RNA-seq	52576	47	35	Aguirre et al.[4]
CPTAC	RNA-seq	28057	124	46	Cao et al.[5]
Dijk	RNA-seq	49677	90	81	Dijk et al.[6]
Moffitt	Microarray	19749	123	83	Moffitt et al.[7]
PACA_AU	Microarray	47265	63	38	Bailey et al.[8]
Puleo	Microarray	20087	288	181	Puleo et al.[9]
TCGA	RNA-seq	20531	144	75	Raphael et al.[10]

TABLE 1 Summaries of PDAC gene expression datasets. Used in case study prediction model of survival.

The subjects used in the analyses were restricted to those with primary pancreatic ductal adenocarinoma samples. Of the total genes listed, 420 of these genes were both part of the 500 member gene list that Moffitt et al.[7] specified as relevant for classifying subjects into the basal vs classical subtypes and common among all of the datasets. (The 500 member gene list is also provided within the aforementioned GitHub repository). We removed the bottom 20% of these 420 genes by rank-transforming the genes and then removing the genes with the lowest average rank.

Because the datasets are a mix of RNA-seq and microarray datasets, we integrated the data together using the data integration rank transformation technique as specified by Rashid et al.[11] This integration technique creates top scoring pairs (TSPs). To illustrate the interpretation of TSPs, let $g_{ki,A}$ and $g_{ki,B}$ be the raw expression of genes A and B in subject i of group k. For each gene pair $(g_{ki,A}, g_{ki,B})$, the TSP is an indicator $I(g_{ki,A} > g_{ki,B})$ which specifies which of the two genes has higher expression in the subject. We denote a TSP predictor as “GeneA_GeneB”.

To pick the TSPs used in our analyses, we first removed TSPs that had low variation across the samples. If the average value of the TSP was less than 0.10 or greater than 0.90, they were not included in later analyses. Next, we fit single covariate proportional hazard mixed effects models using the **coxme** R package,[12] where the models contained the TSP, a random intercept, and a random slope for the TSP. From these models, we extracted the estimated log-likelihoods and sorted the TSPs based on these log-likelihoods (highest to lowest).

After sorting the TSPs by their log-likelihoods, we removed TSPs with high correlation. Starting with the TSP with the maximum log-likelihood, we removed all other TSPs with lower (i.e. more negative) log-likelihood values that shared one of the genes in the first TSP. Then we moved to the next TSP with the second highest log-likelihood and repeated this process, continuing in this way for all downstream TSPs. At the end of this procedure, we were left with 168 TSPs.

2.2 | Case study: Sensitivity analyses

We performed sensitivity analyses on our case study by running the Elastic Net variable selection procedure with alternative values of π—the value that represents the balance between ridge regression and the MCP penalty ($\pi = 0$ represents ridge regression, $\pi = 1$ represents the MCP penalty)—and alternative values for the number of latent common factors r (for the **phmmPen_FA** procedure). We considered values of $\pi = 0.7, 0.8, 0.9, 1.0$.

In addition to estimating the number of latent common factors using the Growth Ratio procedure, which estimated a value of $r = 2$, we also fit the model assuming $r = 3$ because the simulations given in the main paper indicated that the Growth Ratio method may underestimate r.

The times to complete the **phmmPen_FA** variable selection procedure for $\pi = \{0.7, 0.8, 0.9, 1.0\}$ was between 1.4 and 2.5 hours when $r = 2$ and between 2.1 and 3.1 hours when $r = 3$.

For readers interested in more details about the results from these sensitivity analyses, please see the GitHub repository https://github.com/hheiling/paper_phmmPen_FA and explore the “Replication/replication_case_study_sensitivity.R” code, which examines the PDAC case study results saved within the “Replication/Paper_Results/PDAC_Selection_Results.RData” object.

3 | SUPPLEMENTARY MATERIAL: ADDITIONAL SIMULATIONS

3.1 | Variable selection assuming naive model with fixed effects only: ncvreg::cv.ncvsurv

At the time of the writing of this paper, there were no survival mixed effects variable selection procedures that simultaneously selected both fixed and random effects and could handle the dimensions used on our simulations and case study. Therefore, we did not have a method with which we could directly compare the performance of our **phmmPen_FA** method. Consequently, we decided to compare our method with a 'naive' variable selection method that does not account for random effects. If so inclined, a researcher interested in selecting predictors from a high dimensional survival dataset without using our **phmmPen_FA** method could use such a 'naive' variable selection approach and assume that the selected predictors represented the true non-zero fixed and random effects. We chose to compare our method with the **ncvreg** R package,[2] which has a `cv.ncvsurv` function that can perform variable selection on survival data using the MCP penalty and select the 'best' penalty parameter and corresponding model using k-fold cross validation.

We simulated $p = 100$ piecewise constant hazard data using the same procedure and the same parameter conditions (including fixed effect coefficient values, number of groups, relative size of the $\boldsymbol{B}$ matrix, and true $r = 3$) as used in Section 3 of the main manuscript. We fit this data using the `cv.ncvsurv` function of the **ncvreg** R package; we specified an MCP penalty and a minimum penalty value as $0.05\lambda_{max}$ (similar to how we specified the minimum penalty in our variable selection procedure), but all other possible input arguments were set as the default `cv.ncvsurv` argument options to mimic how a typical researcher would use the package.

As we can see from the true positive rates presented in Table 2, if we treated the selected predictors as the true non-zero fixed and random effects, we would often significantly underestimate the number of true fixed and random effects (if we assumed that the selected predictors represented both the non-zero fixed and non-zero random effects). In many of the cases considered in our simulations, the true positive rates from the **ncvreg** procedure were around 10% less than the true positive rates seen in Table 1 of the main paper (for both fixed and random effects). The **ncvreg** procedure only performed comparably well with our method when the true fixed effect coefficient values were large ($\beta = 1.0$, corresponding to a hazard ratio of 2.72) and the true random effect variance was relatively small.

Of course, it should be acknowledged that the **ncvreg** R package method performs much faster compared with our method since they do not have to account for random effects in their model. In comparison to our method where the median time to complete the algorithm was on the scale of hours, the median time to complete the **ncvreg** variable selection procedure was on the scale of seconds.

Although the **ncvreg** procedure is much faster, we have demonstrated that our method **phmmPen_FA** outperforms this 'naive' method in many scenarios where the true underlying data contains mixed effects.

β	K	$\boldsymbol{B}$	TP %	FP %	T^{med}	Abs. Dev.
			Fixef	Fixef	(sec)	(Mean)
0.5	5	Small	81.60	2.78	1.33	0.25
		Moderate	79.60	3.54	1.30	0.31
	10	Small	85.20	2.59	1.26	0.29
		Moderate	71.60	2.67	2.26	0.33
1.0	5	Small	97.80	1.53	2.35	0.50
		Moderate	88.80	2.51	2.28	0.61
	10	Small	99.60	1.73	2.39	0.56
		Moderate	95.00	2.67	2.33	0.68

TABLE 2 Variable selection results assuming naive model with fixed effects only, where the simulated piecewise constant hazard mixed effects data ($p = 100$) was fit using the `ncvreg::cv.ncvsurv` function. Results include the true positive (TP) percentages for fixed effects, false positive (FP) percentages for fixed effects, the median time in seconds for the algorithm to complete ($T^{med}(sec)$), and the average of the mean absolute deviation (Abs. Dev. (Mean)) between the coefficient estimates and the true β values across all simulation replicates. Column $\boldsymbol{B}$ describes the general size of both the variances and eigenvalues of the resulting $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T$ random effects covariance matrix.

3.2 Data simulated using Weibull distribution

In the simulations presented in Section 3 of the main paper, we both simulated data from a piecewise constant hazard mixed effects distribution and fit the data assuming a piecewise constant hazard mixed effects model. In this section, we present simulations in which we simulated survival data using a Weibull mixed effects distribution and then performed variable selection on the data using our **phmmPen_FA** method, assuming a piecewise constant hazard mixed effects model.

To simulate the survival data from a Weibull mixed effects distribution, we kept many aspects of the data simulation consistent with the piecewise constant hazards mixed effects simulations described in Section 3 of the main paper. We used the same sample sizes and number of groups, the same number of non-zero fixed and random effects, and the same fixed effect coefficients and random effect coefficients.

We only changed how the survival times were generated. We simluated Weibull survival times in a very similar manner to how data is simulated from the **simsurv** R package.[13] In order to generate survival times from the Weibull distribution, we first sampled survival probabilities from a uniform distribution, $S_{ki} \sim Unif(0, 1)$, for all subjects i in group k. The event times were then calculated as:

$$T_{ki} = \left(\frac{-\log(S_{ki})}{\exp(\psi^{\star} + \eta_{ki})} \right)^{1/\omega}, \tag{6}$$

where $\omega = 5$ is the Weibull shape parameter, $\psi^{\star} = 1.0$ is the log of the Weibull scale parameter, and η_{ki} is the linear predictor for subject i in group k (defined in the same way as in Section 3 of the main paper).

As before, censoring times $\boldsymbol{C} = (\boldsymbol{C}_1^T, ..., \boldsymbol{C}_k^T)^T$ where $\boldsymbol{C}_k = (C_{k1}, ..., C_{kn_k})^T$ were simulated from the uniform distribution $C_{ki} \sim Unif(0, 5)$, which assumes an end to follow-up after 5 years. When we combined these censoring times with the choices of $\omega = 5$ for the Weibull shape parameter and $\psi^{\star} = 1.0$ for the log of the Weibull scale parameter, we achieved censoring rates and median follow-up times very similar between these Weibull mixed effects simulations and the previous piecewise constant hazard mixed effects simulations: the Weibull mixed effects censoring rates across all simulated data fell between 12% and 26%, with average censoring rates across the 100 simulation replicates for each of the simulation conditions between 16% and 20%, and the average of the median follow-up times were approximately 0.66 for the various conditions.

β	K	$\boldsymbol{B}$	Data	Avg. r	TP %	FP %	TP %	FP %	T^{med}	Abs. Dev.	$\lVert\boldsymbol{D}\rVert_F$
			Type		Fixef	Fixef	Ranef	Ranef		(Mean)	
0.5	5	Small	PCH	2.00	90.80	2.46	91.60	1.40	2.34	0.22	0.42
			Weibull	2.00	90.40	2.00	91.40	1.32	1.96	0.20	0.34
		Moderate	PCH	2.00	91.00	4.18	92.40	2.13	3.78	0.32	0.70
			Weibull	2.00	89.00	2.46	93.20	1.81	2.72	0.27	0.69
	10	Small	PCH	2.07	94.00	3.28	95.80	1.63	2.66	0.17	0.31
			Weibull	2.03	94.20	2.88	95.00	1.19	2.66	0.17	0.28
		Moderate	PCH	2.20	86.20	6.42	93.40	3.83	3.02	0.23	0.61
			Weibull	2.04	86.80	3.85	93.60	2.24	2.62	0.21	0.63
1.0	5	Small	PCH	2.00	99.00	1.09	93.20	0.54	3.50	0.25	0.39
			Weibull	2.00	99.60	0.64	91.60	0.43	2.60	0.29	0.31
		Moderate	PCH	2.00	95.20	2.75	94.20	1.22	3.63	0.34	0.71
			Weibull	2.00	95.20	1.18	93.80	0.92	3.34	0.38	0.68
	10	Small	PCH	2.13	99.80	1.20	97.00	0.34	2.97	0.23	0.30
			Weibull	2.06	99.80	0.84	96.20	0.21	2.25	0.31	0.27
		Moderate	PCH	2.19	98.20	4.22	98.80	0.66	4.02	0.33	0.61
			Weibull	2.18	98.20	2.43	97.60	0.52	3.69	0.41	0.63

TABLE 3 Variable selection results comparing two types of simulated survival data using Weibull mixed effects or piecewise constant hazard (PCH) mixed effects distribution assumptions ($p = 100$), including true positive (TP) percentages for fixed and random effects, false positive (FP) percentages for fixed and random effects, the median time in hours for the algorithm to complete (T^{med}), and the average of the mean absolute deviation (Abs. Dev. (Mean)) between the coefficient estimates and the true β values across all simulation replicates. Column $\boldsymbol{B}$ describes the general size of both the variances and eigenvalues of the resulting $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T$ random effects covariance matrix. Column 'Avg. r' reports the average Growth Ratio (GR) estimate of r calculated within the algorithm. Column $\lVert\boldsymbol{D}\rVert_F$ represents the average across simulation replicates of the Frobenius norm of the difference ($\boldsymbol{D}$) between the estimated random effects covariance matrix $\hat{\boldsymbol{\Sigma}}$ and the true random effects covariance matrix $\boldsymbol{\Sigma}$; the Frobenius norm was standardized by the number of true random effects selected in the model.

Table 3 shows the **phmmPen_FA** output results for both the Weibull-simulated data and the piecewise constant hazard (PCH)-simulated data (the PCH simulated data was copied from Table 1 of the main paper for ease of comparison). The true positive rates for the fixed and random effects are very similar between the two data generation mechanisms for all simulation conditions considered. The false positive rates for the fixed and random effects are often a bit better (i.e. smaller) for the Weibull-generated data. The bias for the fixed and random effects coefficients are also often slightly better (slightly smaller) for the Weibull-generated data. In summary, the **phmmPen_FA** method performs well on survival data regardless of whether it is applied to piecewise constant hazard-generated data or Weibull-generated data.

The only measure in which the piecewise constant hazard-generated data had a clearly better performance than the Weibull-generated data was in the estimation of the number of latent factors, r. However, the greater mis-specification of r in the Weibull-generated data did not appear to negatively impact the variable selection performance or coefficient bias.

3.3 | Data simulated with unequal fixed and random effects

In the simulations presented in Section 3 of the main paper, we assumed that the true fixed effects predictors (i.e. predictors with corresponding non-zero fixed effects coefficients) were equal to the true random effects predictors (i.e. predictors with corresponding non-zero random effects coefficients). In this section, we present simulations where we tested the performance of our **phmmPen_FA** procedure when the number of true fixed effects predictors was not the same as the number of true random effects predictors.

We simulated piecewise constant hazard mixed effects survival data in a similar manner to the simulations discussed in Section 3 of the main paper. However, in these simulations, we increased the number of true fixed effects predictors to $p^* = 10$ and kept the number of true random effects the same as the original simulations ($q^* = 6$ representing 5 random effect predictors and a random intercept), where the true random effect predictors were a subset of the true fixed effects predictors. For brevity, we restricted the simulation conditions to those with the true fixed effects coefficients of $\beta = 1.0$. All simulations estimated the number of latent factors r using the Growth Ratio procedure (GR).

From Table 4, we see that the selection of the true random effect predictors were unaffected by having unequal numbers of true fixed and random effects predictors since the random effect true positive rates and false positive rates are very similar to the random effect true and false positive rates presented in Table 1 of the main paper for corresponding simulation set-ups. We conclude that our **phmmPen_FA** method can accurately select fixed and random effects predictors even when the true fixed effects are not exactly equal to the true random effects.

β	K	$\boldsymbol{B}$	Avg. r	TP % Fixef	FP % Fixef	TP % Ranef	FP % Ranef	T^{med}	Abs. Dev. (Mean)	$\|\|\boldsymbol{D}\|\|_F$
1.0	5	Small	2.00	99.00	1.09	93.20	0.54	3.50	0.25	0.39
		Moderate	2.00	95.20	2.75	94.20	1.22	3.63	0.34	0.71
	10	Small	2.15	99.80	1.20	97.00	0.34	2.97	0.23	0.30
		Moderate	2.21	98.20	4.22	98.80	0.66	4.02	0.33	0.61

TABLE 4 Variable selection results for $p = 100$ piecewise constant hazard mixed effects simulation results when the true number of fixed effects predictors (10) is not equal to the true number of random effects predictors (5, subset of true fixed effects). Results include true positive (TP) percentages for fixed and random effects, false positive (FP) percentages for fixed and random effects, the median time in hours for the algorithm to complete (T^{med}), and the average of the mean absolute deviation (Abs. Dev. (Mean)) between the coefficient estimates and the true β values across all simulation replicates. Column $\boldsymbol{B}$ describes the general size of both the variances and eigenvalues of the resulting $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T$ random effects covariance matrix. Column 'Avg. r' reports the average Growth Ratio (GR) estimate of r calculated within the algorithm. Column $||\boldsymbol{D}||_F$ represents the average across simulation replicates of the Frobenius norm of the difference ($\boldsymbol{D}$) between the estimated random effects covariance matrix $\hat{\boldsymbol{\Sigma}}$ and the true random effects covariance matrix $\boldsymbol{\Sigma}$; the Frobenius norm was standardized by the number of true random effects selected in the model.

3.4 | Purposeful underestimation of latent factors r

In Table 1 of the main paper, we found that even when the Growth Ratio (GR) procedure underestimated the true number of latent factors r in the underlying model, the variable selection performance and the coefficient bias were no more than minimally affected. However, since the true r value of 3 and the commonly estimated r value of 2 were not far apart, we wanted to test the performance of **phmmPen_FA** when we purposefully underestimate r to values further from the truth.

The purpose of the simulations presented in this section is to determine whether it is appropriate to purposefully specify a value of r that likely underestimates the true r. Because the speed of the **phmmPen_FA** algorithm depends on the size of the latent space (i.e. the value of r) and the speed decreases with larger values of r, one could consider purposefully underestimating r in order to help speed up the **phmmPen_FA** procedure. We aim to determine if an improvement in speed of the algorithm comes with costs related to the variable selection accuracy and/or the size of the coefficient biases.

We simulated piecewise constant hazard mixed effects survival data in a similar manner to the procedure outlined in Section 3 of the main paper, except we simulated the true number of latent factors r as a value of 5. The new $\boldsymbol{B}$ matrix, now with 5 columns, is given in Section 1.1 of this supplement. For brevity, we restricted our consideration to cases when the number of groups $K = 10$ and the relative size of $\boldsymbol{B}$ is moderate (the corresponding variances of the random effect covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T$ associated with this new $\boldsymbol{B}$ are similar to the variances corresponding to the $\boldsymbol{B}_{moderate}$ described in (1)) because in Table 1 of the main paper, the Growth Ratio (GR) procedure was the most accurate in estimating r under these simulated data conditions. Therefore, under these data conditions, the user has a choice of using a reasonable estimate of r from the GR procedure or purposefully underestimating r.

Table 5 shows that, as expected, decreasing r helps decrease the time needed to complete the **phmmPen_FA** procedure. When the true value of the fixed effects coefficient β is relatively large (value 1.0), decreasing r has a greater impact on the reduction in time. As r decreases and becomes further from the true value of 5, we see the following patterns: the true positives of the fixed and random effects generally decreases; the false positives of the random effects generally increases; and the bias of the random effects covariance matrix (as represented by the Frobenius norm results) generally increases. However, the differences between variable selection and bias performances across values of r are minor. As expected, the differences between $r = 5$ vs $r = 2$ are more pronounced than the differences between $r = 5$ and $r = 3$.

In general, these authors still recommend using a reasonable value of r from the Growth Ratio procedure in the **phmmPen_FA** procedure. However, interested users can determine for themselves whether a reduction in time is worth the cost of the observed performance reduction.

β	K	$\boldsymbol{B}$	r used	TP %	FP %	TP %	FP %	T^{med}	Abs. Dev.	$\lVert\boldsymbol{D}\rVert_F$
				Fixef	Fixef	Ranef	Ranef		(Mean)	
0.5	10	Moderate	5	86.80	9.40	93.80	2.98	6.33	0.29	0.74
			3	84.60	8.09	93.80	3.18	4.00	0.25	0.81
			2	81.00	7.76	91.80	3.33	3.66	0.24	0.89
1.0	10	Moderate	5	97.40	4.60	97.20	0.27	13.75	0.31	0.72
			3	97.20	6.20	97.00	0.75	8.33	0.35	0.80
			2	97.40	5.57	93.60	1.27	4.11	0.43	0.89

TABLE 5 Variable selection results for $p = 100$ piecewise constant hazard mixed effects simulations assuming the true r of 5 vs purposeful underestimation of r using values of 3 or 2. Results include true positive (TP) percentages for fixed and random effects, false positive (FP) percentages for fixed and random effects, the median time in hours for the algorithm to complete (T^{med}), and the average of the mean absolute deviation (Abs. Dev. (Mean)) between the coefficient estimates and the true β values across all simulation replicates. Column $\boldsymbol{B}$ describes the general size of both the variances and eigenvalues of the resulting $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T$ random effects covariance matrix. Column 'Avg. r' reports the average Growth Ratio (GR) estimate of r calculated within the algorithm. Column $\lVert\boldsymbol{D}\rVert_F$ represents the average across simulation replicates of the Frobenius norm of the difference ($\boldsymbol{D}$) between the estimated random effects covariance matrix $\hat{\boldsymbol{\Sigma}}$ and the true random effects covariance matrix $\boldsymbol{\Sigma}$; the Frobenius norm was standardized by the number of true random effects selected in the model.

3.5 | Alternative number of time intervals J

In Section 3 of the main paper, we discuss our reasoning for our choice of assuming $J = 8$ time intervals within the piecewise constant hazard mixed effects model assumptions within the **phmmPen_FA** procedure. In this section, we examine the performance of the **phmmPen_FA** procedure when we assume different numbers of J time intervals.

We simulate the data the same way as we specified in Section 3 of the main paper. For brevity, we restrict our consideration to the data conditions with fixed effect coefficient values of $\beta = 0.5$. When performing the **phmmPen_FA** variable selection procedure, we specified a range the number of time intervals from 5 to 10. We chose this range because our review of the literature suggests that it is common to choose between 5 and 10 time intervals for piecewise constant hazard models when they are fitting real data.[14,15,16,17,18] All simulations used the Growth Ratio estimates for the number of latent factors r.

In general, there is very little difference between the variable selection and bias results presented here in Table 6 for several values of J compared with the results presented in Table 1 of the main paper when $J = 8$. As the number of time intervals J increases, the fixed and random effect true positive rates increase slightly from 5 to 9 but level off or decrease slightly when $J = 10$. The impact that J has on the false positive rates for the fixed and random effects varies depending on other data conditions. As J increases, the random effect coefficient bias tends to decrease slightly (see the Frobenius norm values), while the fixed effect coefficient bias (the mean absolute deviation values) remains very consistent across values of J. The most obvious impact that the value of J seems to have is on the estimation of r, where smaller J values in the range of 5 to 10 improve the accuracy of the Growth Ratio r estimate when $K = 10$, and the time it takes to complete the **phmmPen_FA** variable selection procedure, where increased values of J increase the median time needed to complete the procedure. See Section 4.1 for a further discussion on how J impacts the Growth Ratio estimation of r.

We conclude from these simulations that although we do not provide a recommendation for selecting the 'best' number of time intervals J to use in the piecewise constant hazard mixed effects model, we think using any value of J in the range of 5 to 10 gives reasonable results.

β	K	$\boldsymbol{B}$	J	Avg. r	TP %	FP %	TP %	FP %	T^{med}	Abs. Dev.	$\|\|\boldsymbol{D}\|\|_F$
					Fixef	Fixef	Ranef	Ranef		(Mean)	
0.5	5	Small	5	2.00	89.60	2.29	88.80	1.12	1.83	0.21	0.46
			6	2.00	90.20	2.63	90.80	1.23	1.56	0.22	0.46
			9	2.00	91.80	2.65	92.20	1.34	2.05	0.22	0.40
			10	2.00	92.40	2.32	92.40	1.79	2.42	0.22	0.38
		Moderate	5	2.00	88.00	3.62	90.80	2.15	2.30	0.30	0.79
			6	2.00	90.00	3.14	93.40	1.80	3.07	0.30	0.75
			9	2.00	90.20	3.87	94.20	1.83	3.80	0.31	0.71
			10	2.00	89.80	4.33	94.00	2.34	5.72	0.30	0.68
	10	Small	5	2.34	93.40	2.67	95.80	0.88	1.49	0.17	0.34
			6	2.21	94.00	2.83	94.80	1.12	2.19	0.17	0.34
			9	2.01	95.20	2.84	96.20	2.34	3.32	0.17	0.30
			10	2.02	94.60	2.72	95.80	2.37	3.80	0.16	0.29
		Moderate	5	2.46	87.60	8.95	94.40	3.28	2.04	0.22	0.65
			6	2.43	85.80	5.58	93.80	3.19	2.57	0.22	0.62
			9	2.21	87.80	7.42	93.40	4.82	3.40	0.23	0.61
			10	2.04	87.20	6.61	93.00	4.78	4.23	0.22	0.60

TABLE 6 Variable selection results assuming a range of alternative number of time intervals J in the $p = 100$ piecewise constant hazard mixed effects simulations. Results include true positive (TP) percentages for fixed and random effects, false positive (FP) percentages for fixed and random effects, the median time in hours for the algorithm to complete (T^{med}), and the average of the mean absolute deviation (Abs. Dev. (Mean)) between the coefficient estimates and the true β values across all simulation replicates. Column $\boldsymbol{B}$ describes the general size of both the variances and eigenvalues of the resulting $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T$ random effects covariance matrix. Column 'Avg. r' reports the average Growth Ratio (GR) estimate of r calculated within the algorithm. Column $||\boldsymbol{D}||_F$ represents the average across simulation replicates of the Frobenius norm of the difference ($\boldsymbol{D}$) between the estimated random effects covariance matrix $\hat{\boldsymbol{\Sigma}}$ and the true random effects covariance matrix $\boldsymbol{\Sigma}$; the Frobenius norm was standardized by the number of true random effects selected in the model.

4 | SUPPLEMENTARY MATERIAL: FURTHER DISCUSSION

4.1 | Growth Ratio procedure experimentation

In general, the accuracy of the Growth Ratio procedure is impacted by several qualities about the data. In Heiling et al.[19] simulations on binary outcome data, it was found that the Growth Ratio procedure has improved accuracy for estimating the number of latent factors r under the following conditions:

- When the number of groups increases
- When the number of observations per group increases
- When the ratio of number of random effect features to number of observations per group decreases
- When the relative "size" of the $\boldsymbol{B}$ matrix increases, i.e. when the eigenvalues of the matrix $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T$ increases

Under non-ideal circumstances that represent the opposite of the above conditions (e.g. when the number of groups decreases instead of increases), the Growth Ratio procedure tends to underestimate the number of latent factors r.

In the survival outcome simulations presented in Section 3 of the main manuscript, we see that the Growth Ratio procedure tended to consistently underestimate the value of r, and resulted in greater underestimation compared to the results observed by Heiling et al.[19] This can be explained by several parameters used within the simulations of Section 3, including a relatively smaller number of simulated groups K (5 and 10 in this manuscript, compared to 25 in Heiling et al.[19]) and more moderate "sizes" of $\boldsymbol{B}$ used.

We also saw in Section 3.5 of this supplement that as we decreased the number of time intervals J, we could further improve the accuracy of the Growth Ratio procedure under some select simulation conditions while keeping the number of groups and the "size" of $\boldsymbol{B}$ the same. When we define the interval boundaries for a certain pre-specified number of time intervals J, we design each interval to have an approximately equal number of events per interval. However, we do not make an effort to adjust the interval boundaries so that each group within the data has some minimum number of events per interval. This could impact the Growth Ratio procedure because this procedure fits and later utilizes group-specific coefficient estimates (which create pseudo random effects); if a group has too few events within an interval, this could reduce the stability of the group-specific coefficient estimates. Consequently, this could negatively impact the quality of the pseudo random effect estimates used within the Growth Ratio procedure. If this hypothesis is true, we would expect that the Growth Ratio procedure would see improved estimation when the number of time intervals within the piecewise constant hazard survival mixed effects model decreases, since this would increase the likelihood that an adequate number of events per group is observed within each time interval. This improvement may be offset if too few time intervals are used, which could itself lead to greater bias in the group-specific coefficients used to calculate the pseudo random effects.

To further test this hypothesis, we tested the performance of the Growth Ratio procedure assuming a different number of time intervals J within the piecewise constant hazard survival mixed effects model for a subset of the simulation conditions used within Section 3 of the main paper. We compared the average value of r estimated from 100 simulation replicates assuming $J = \{8, 5, 3\}$, where $J = 8$ was used within the original main manuscript simulations. The data was simulated the same way as described in Section 3 of the main manuscript, assuming the fixed effect coefficient values of $\beta = 0.5, 1.0$, "moderate" $\boldsymbol{B}$, the number of groups $K = 5, 10$, and the true number of latent factors r is equal to 3. Table 7 shows the results of this experimentation.

β	K	$\boldsymbol{B}$	Avg. r	Avg. r	Avg. r
			$J = 8$	$J = 5$	$J = 3$
0.5	5	Moderate	2.00	2.00	2.00
	10	Moderate	2.20	2.46	2.32
1.0	5	Moderate	2.00	2.00	2.00
	10	Moderate	2.19	2.49	2.37

TABLE 7 The average estimate of the number of latent factors r produced by the Growth Ratio procedure when the piecewise constant hazard survival mixed effects model assumed $J = 8, 5, 3$ number of time intervals. The true value of r in the simulations was 3. Column $\boldsymbol{B}$ describes the general size of both the variances and eigenvalues of the resulting $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T$ random effects covariance matrix.

We see from the results in Table 7 that when the number of groups K in the data is 10, decreasing the number of time intervals from $J = 8$ to 5 does improve the Growth Ratio estimation procedure (i.e. the average value of r across the simulation replicates is closer to the true value of 3). However, further decreasing the number of time intervals to $J = 3$ did not further improve the estimation procedure compared with $J = 5$. When the number of groups K in the data is 5, we observe no change in the average estimate of r regardless of the value of J.

These results support our earlier hypothesis. Based on the results in this section and the results presented in Section 3.5 of this supplement, we conclude that decreasing J up to a certain point does improve the Growth Ratio estimate of r, but decreasing J too far has a negative impact on this estimation of r. We also saw that decreasing J has some slightly negative impacts on the variable selection accuracy and random effect coefficient bias (see Section 3.5 for details), so decreasing J has some other costs to this **phmmPen_FA** procedure.

In order to improve the estimation of r within **phmmPen_FA**, we could have potentially modified the code so that the Growth Ratio estimation procedure assumed a different number of time intervals compared to the later variable selection procedure used within **phmmPen_FA**. However, we saw from the simulation results presented in Table 1 of the main manuscript that compared with using the true value of r, underestimating r did not significantly impact the true and false positive rates of the variable selection procedure, nor did it significantly impact the bias of the fixed effect coefficients. Therefore, we decided that further complicating our **phmmPen_FA** procedure with further modification options for the Growth Ratio procedure was not a valuable contribution.

SUPPORTING INFORMATION

Software in the form of R code for the **phmmPen_FA** procedure is available through the **glmmPen** package in CRAN https://cran.r-project.org/package=glmmPen and the GitHub repository https://github.com/hheiling/glmmPen. The **phmmPen_FA** procedure is implemented through the `phmmPen_FA()` function within this **glmmPen** R package. Code to replicate the simulation and the case study analyses and results is available through the GitHub repository https://github.com/hheiling/paper_phmmPen_FA.

REFERENCES

1. Ibrahim JG, Zhu H, Garcia RI, Guo R. Fixed and random effects selection in mixed effects models. *Biometrics.* 2011;67(2):495–503.
2. Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics.* 2011;5(1):232–253.
3. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software.* 2010;33(1):1–22.
4. Aguirre AJ, Nowak JA, Camarda ND, et al. Real-time genomic characterization of advanced pancreatic cancer to enable precision medicine. *Cancer discovery.* 2018;8(9):1096–1111.
5. Cao L, Huang C, Zhou DC, et al. Proteogenomic characterization of pancreatic ductal adenocarcinoma. *Cell.* 2021;184(19):5031–5052.
6. Dijk F, Veenstra VL, Soer EC, et al. Unsupervised class discovery in pancreatic ductal adenocarcinoma reveals cell-intrinsic mesenchymal features and high concordance between existing classification systems. *Scientific reports.* 2020;10(1):337.
7. Moffitt RA, Marayati R, Flate EL, et al. Virtual microdissection identifies distinct tumor-and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature genetics.* 2015;47(10):1168.
8. Bailey P, Chang DK, Nones K, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature.* 2016;531(7592):47–52.
9. Puleo F, Nicolle R, Blum Y, et al. Stratification of pancreatic ductal adenocarcinomas based on tumor and microenvironment features. *Gastroenterology.* 2018;155(6):1999–2013.
10. Raphael BJ, Hruban RH, Aguirre AJ, et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer cell.* 2017;32(2):185–203.
11. Rashid NU, Li Q, Yeh JJ, Ibrahim JG. Modeling between-study heterogeneity for improved replicability in gene signature selection and clinical prediction. *Journal of the American Statistical Association.* 2020;115(531):1125–1138.
12. Therneau TM. *coxme: Mixed Effects Cox Models*. 2022. R package version 2.2-18.1.
13. Brilleman SL, Wolfe R, Moreno-Betancur M, Crowther MJ. Simulating Survival Data Using the simsurv R Package. *Journal of Statistical Software.* 2020;97(3):1–27. doi: 10.18637/jss.v097.i03
14. Austin PC. A tutorial on multilevel survival analysis: methods, models and applications. *International Statistical Review.* 2017;85(2):185–203.
15. SAS Institute Inc. . *SAS/STAT Software, Version 9.4 15.1 User's Guide*. Cary, NC: 2018.
16. Li Y, Gail MH, Preston DL, Graubard BI, Lubin JH. Piecewise exponential survival times and analysis of case-cohort data. *Statistics in medicine.* 2012;31(13):1361–1368.
17. Qiou Z, Ravishanker N, Dey DK. Multivariate survival analysis with positive stable frailties. *Biometrics.* 1999;55(2):637–644.
18. Friedman M. Piecewise exponential models for survival data with covariates. *The Annals of Statistics.* 1982;10(1):101–113.
19. Heiling HM, Rashid NU, Li Q, Peng XL, Yeh JJ, Ibrahim JG. Efficient computation of high-dimensional penalized generalized linear mixed models by latent factor modeling of the random effects. *Biometrics.* 2024;80(1):ujae016.