DOI: 000

Supplementary Materials for 'Efficient Computation of High-Dimensional Penalized Generalized Linear Mixed Models by Latent Factor Modeling of the Random Effects'

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

[1]Department of Biostatistics, University of North Carolina Chapel Hill, Chapel Hill, NC
[2]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC
[3]Department of Surgery, University of North Carolina Chapel Hill, Chapel Hill, NC
[4]Department of Pharmacology, University of North Carolina Chapel Hill, Chapel Hill, NC
[*]email: hmheiling@gmail.com

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

1. Web Appendix: Supplementary glmmPen_FA procedure details

This section of the Web Appendix provides additional details about several aspects of the **glmmPen_FA** procedure, including details about the M-step and MCECM algorithms, model selection, tuning parameter selection, and initialization and convergence.

1.1 *Method algorithms*

Algorithms 1 and 2 provide additional details about the M-step and overall MCECM procedure within **glmmPen_FA**.

Let s represent the iteration of the MCECM algorithm, and let h represent the iteration within a particular M-step of the MCECM algorithm.

Algorithm 1 M-step of the MCECM algorithm

1. Coefficient parameter estimates from the previous M-step, $\boldsymbol{\theta}^{(s-1)}$, are used to initialize the coefficient parameters of the current M-step at M-step iteration $h = 0$, denoted $\boldsymbol{\theta}^{(s,0)}$.
2. Conditional on $\boldsymbol{b}^{(s,h-1)}$ and $\tau^{(s-1)}$, each $\beta_j^{(s,h)}$ for $j = 1, ..., p$ is given a single update using the Majorization-Minimization algorithm specified by Breheny and Huang (2015).
3. For each group k in $k = 1, ..., K$, the augmented matrix $\tilde{\boldsymbol{z}}_{ki} = (\tilde{\boldsymbol{\alpha}}_k^{(s)} \otimes \boldsymbol{z}_{ki})J$ is created for $i = 1, ..., n_k$ where $\tilde{\boldsymbol{\alpha}}_k^{(s)} = ((\boldsymbol{\alpha}_k^{(s,1)})^T, ..., (\boldsymbol{\alpha}_k^{(s,M)})^T)^T$. This augmented matrix is used in the random effect portion of the linear predictor specified in Equation 4 of the main manuscript. This augmented matrix is used to calculate the terms needed to update random effect coefficients $\boldsymbol{b}_t^{(s,h)}$ for $t = 1, ..., q$, see Equation 2.9 in Breheny and Huang (2015).
4. Conditional on the $\tau^{(s-1)}$ and the recently updated $\boldsymbol{\beta}^{(s,h+1)}$, each $\boldsymbol{b}_t^{(s,h)}$ for $t = 1, ..., q$ is updated using the Majorization-Minimzation coordinate descent grouped variable selection algorithm specified by Breheny and Huang (2015).
5. Steps 2 through 4 are repeated until the M-step convergence criteria are reached or until the M-step reaches its maximum number of iterations.
6. Conditioning on the newly updated $\boldsymbol{\beta}^{(s)}$ and $\boldsymbol{b}^{(s)}$, $\tau^{(s)}$ is updated (generically, using the Newton-Raphson algorithm; for the Gaussian family, using a quantity derived from the Q-function approximation in Equation 9).

Algorithm 2 Full MCECM algorithm for single (λ_0, λ_1) penalty combination

1. Fixed and random effects $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{b}^{(0)}$ are initialized as discussed in Web Appendix Section 1.4.
2. E-step: In each E-step for EM iteration s, a burn-in sample from the posterior distribution of the random effects is run and discarded. A sample of size $M^{(s)}$ from the posterior is then drawn and retained for the M-step. (See Web Appendix Section 1.4 for details).
3. M-step: Parameter estimates of $\boldsymbol{\beta}^{(s)}$, $\boldsymbol{b}^{(s)}$, and $\tau^{(s)}$ are then updated as described in the M-step procedure given above.
4. Steps 2 and 3 are repeated until the the average Euclidean distance between the current coefficient vector $(\boldsymbol{\beta}^{(s)T}, \boldsymbol{b}^{(s)T})^T$ and the coefficient vector from $t = 2$ EM iterations back is less than a pre-specified threshold of 0.0015 for two consecutive EM iterations or until the maximum number of EM iterations (25) is reached. (See Web Appendix Section 1.4 for further convergence details).

1.2 *Model selection*

This section provides details on how the **glmmPen_FA** algorithm selects the optimal tuning parameter combination. In all simulations and analyses discussed in this paper, we set the candidate random effects predictors equal to the candidate fixed effects predictors (i.e. $q = p$) and aim to have the algorithm select the true fixed and random effect predictors.

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 (Ibrahim et al., 2011). 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 Web Appendix Section 2 for performance results). The **glmmPen** method uses the same two-stage model selection approach within its model selection procedure.

Before we run the above described two-stage model selection procedure, we run a pre-screening step. This pre-screening step is used in both the **glmmPen_FA** and **glmmPen** methods. This step was designed to filter out a few random effects before running the main algorithm. The pre-screening step fits a minimum penalty model using relatively lax convergence criteria in comparison to the **glmmPen** package's default convergence criteria; if a random effect has been penalized out of the model at the end of the pre-screening step or if the variance estimate for a predictor is below a threshold level value 0.01, that random effect is removed from consideration for the rest of the model selection procedure. The goal of this pre-screening step is to help reduce the time needed to run the remainder of the model selection procedure by reducing the number of random effects considered in the algorithm.

In addition to helping to speed up the algorithm by reducing the number of random effects considered in the algorithm, this pre-screening step also helps to improve the convergence of the minimum penalty model used to calculate the MCMC posterior draws used within the BIC-ICQ calculation. The coefficient estimates from the end of the pre-screening procedure are used to initialize the coefficients for this minimum penalty model. In essence, this increases the number of EM iterations for the minimum penalty model, thereby hopefully improving the model fit of the minimum penalty model and the resulting effectiveness of the MCMC posterior draws taken from this model and used in the BIC-ICQ calculations.

Note: The convergence criteria used in the pre-screening step matches the convergence criteria used in the remainder of the model selection algorithm, see Web Appendix Section 1.4 and 1.5 for details.

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 (Breheny and Huang, 2011) to calculate this value.

For all Binomial outcome variable selection simulations where the total number of predictors was 100 and for all case study analyses, we used the following sequence of penalties for both the fixed effects and the rows of the $\boldsymbol{B}$ matrix (or rows of the $\boldsymbol{\Gamma}$ matrix when using **glmmPen**): 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 Binomial outcome variable selection simulations where the total number of predictors was 500, we used the following sequence of penalties: a sequence of 10 penalties from $0.15\lambda_{max}$ 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. In simulations not shown here, using a consistent sequence of 10 penalties from $0.10\lambda_{max}$ to λ_{max} for both sets of parameters resulted in very similar final results, but the variable selection procedure took more time to complete; using a consistent sequence of 10 penalties from $0.15\lambda_{max}$ to λ_{max} for both sets of parameters decreased the random effect true positive results.

For all Binomial outcome variable selection simulations where the total number of predictors was 25, we used the following sequence of penalties for both the fixed effects and the rows of the $\boldsymbol{B}$ matrix (**glmmPen_FA** procedure) or the rows of the $\boldsymbol{\Gamma}$ matrix (**glmmPen** procedure): a sequence of 10 penalties from $0.01\lambda_{max}$ to λ_{max}, with penalty values equidistant from each other on the log scale.

For the Poisson outcome variable selection simulations, a penalty sequence with larger values was needed for both the fixed effects and rows of the $\boldsymbol{B}$ matrix due to the nature of how the data was simulated and fit. In these simulations, the covariate values $x_{ki,j}$ were simulated from a $N(0, \sigma = 0.10)$ distribution for $j = 1, ..., p$ and left unstandardized in the algorithm, whereas in the binomial simulations, the covariate values were simulated from the standard normal distribution $N(0, 1)$ and then standardized so that $\sum_{k=1}^{K} \sum_{i \in n_k} x_{ki,j} = 0$ and $\boldsymbol{x}_j^T \boldsymbol{x}_j / N = 1$ for each j. The fixed effects penalty sequence included $0.30\lambda_{max}$ and $(\delta_{0,1}, ..., \delta_{0,12}) * \lambda_{max}$, where $\delta_{0,i} = 2 + (i - 1)$. The random effect penalty sequence applied to rows of the $\boldsymbol{B}$ matrix included $0.30\lambda_{max}$ and $(\delta_{1,1}, ..., \delta_{1,11}) * \lambda_{max}$, where $\delta_{1,i} = 0.5 + (i - 1)$.

1.4 *Initialization and convergence - glmmPen_FA*

The fixed effects $\boldsymbol{\beta}^{(0)}$ and random effects covariance terms $\boldsymbol{b}^{(0)}$ are initialized at iteration $s = 0$ in one of two ways. We discuss first the initialization procedure used when the package **glmmPen_FA** is used to fit the first model in the sequence of models fit for variable selection. In this scenario, the fixed effects $\boldsymbol{\beta}^{(0)}$ are initialized by fitting a 'naive' model, where we assume no random effects, i.e. all observations are assumed to be independent and identically distributed. This naive model is fit using the coordinate descent techniques of Breheny and Huang (2011).

Based on the initialized fixed effects $\boldsymbol{\beta}^{(0)}$, the predictors initialized with non-zero 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 $\sqrt{0.10/r}$, 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 0.10.

The E-step MCMC chain of the sample of the posterior density $\phi(\boldsymbol{\alpha}_k | \boldsymbol{d}_{k,o}; \boldsymbol{\theta}^{(s)})$ for groups $k = \{1, ..., K\}$ is initialized in iteration $s = 0$ with random draws from the standard normal distribution. For all following iterations $s \geqslant 1$, 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. At each iteration s, a burn-in sample of size 100 is drawn and discarded before the $M^{(s)}$ samples are drawn and kept for use in the next M-step.

When the algorithm performs variable selection, we initialize models with 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.

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

$$||(\boldsymbol{\beta}^{(s)T}, \boldsymbol{b}^{(s)T})^T - (\boldsymbol{\beta}^{(s-t)T}, \boldsymbol{b}^{(s-t)T})^T||_2^2 / d_n^{s-t} < \epsilon_{EM} \tag{1}$$

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

$$\max_j |\beta_j^{(s,f+1)} - \beta_j^{(s,f)}| \cap \max_{t,h} |b_{th}^{(s,f+1)} - b_{th}^{(s,f)}| < \epsilon_m, \tag{2}$$

where b_{th} for $h=1,...,r$ is an individual element of $\boldsymbol{b}_t$, which is the t-th row of the $\boldsymbol{B}$ matrix. The value of ϵ_m was set to 0.001.

1.5 *Initialization and convergence - glmmPen*

The initialization and convergence of **glmmPen** in these simulations and analyses were very similar to the initialization and convergence of **glmmPen_FA**, except we replace the notation for the $\boldsymbol{B}$ matrix with the notation for the $\boldsymbol{\Gamma}$ matrix and we have a different initialization procedure for the random effect covariance matrix. This starting variance is initialized in an automated fashion. First, a GLMM composed of only a fixed and random intercept is fit using the **lme4** package. The random intercept variance from this model is then multiplied by 2, and this value is set as the starting values of the diagonal of the random effects covariance matrix. Similar to the **glmmPen_FA** random effect intialization, the predictors are initialized to have non-zero random effects if they are also initialized to have nonzero fixed effects; otherwise, predictors are initialized to have no random effects.

1.6 *Further clarification of p, q, and r*

We provide here some additional clarifications and discussion about the notation of p, q, and r used in the main paper and this supplementary material. We let the values of p and q represent the full set of candidate predictors for the fixed and random effects, respectively. In other words, these values of p and q refer to the input predictors of the **glmmPen_FA** variable selection procedure. We specify here alternative notations of the true number of non-zero fixed and random effect predictors in the model, $p^* \leqslant p$ and $q^* \leqslant q$, respectively.

Our assumption in the manuscript is that $r \ll q$ (i.e. r is much less than q), meaning the number of latent factors used in the model is much smaller than the total number of candidate random effect predictors considered in the model. In contrast, we do not necessarily assume that the number of latent factors r is much less than the true number of random effects q^*. In other words, we do not assume $r \ll q^*$, although we do assume $r < q^*$.

2. Web Appendix: Supplementary simulation details and results

This section of the Web Appendix describes additional details about the $\boldsymbol{B}$ matrices used within the simulations run in Section 3 of the main manuscript as well as additional simulation results not presented in the main manuscript due to space considerations. All simulation conditions used 100 replicates.

These additional simulations include:

- Comparison between **glmmPen_FA** and **glmmPen**
- Variable selection in Binomial data with $p = 500$ predictors
- Variable selection using the Elastic Net penalty applied to correlated predictors
- Variable selection with alternative β and $\boldsymbol{B}$ matrix sizes
- Variable selection with alternative sample size and number of groups
- Variable selection with alternative number of true random effects
- Variable selection in Poisson data with $p = 100$ predictors

For readers who are interested in other measures from the simulations not reported within the main paper or this Web Appendix, such as the range or standard deviations of the time to complete the simulation replicates, we note here that important output from each simulation replicate is stored in the GitHub repository `https://github.com/hheiling/paper_glmmPen_FA` as well as in supplemental material provided with this paper at the Biometrics website on Oxford Academic. See the "Replication/Paper_Results_Revision" folder with RData files corresponding to each simulation discussed in the main paper and this Web Appendix. The simulation output stored in these RData objects could be used to calculate other measures of interest.

2.1 *B matrices used in simulations*

The transpose of the first 11 rows of the deterministic 'large' $\boldsymbol{B}$ matrices used in the Binomial simulations in Section 3 of the main manuscript are given in the Web Appendix equations (3) and (4), corresponding to $r = \{3, 5\}$, respectively. The deterministic 'moderate' $\boldsymbol{B}$ matrices are these large $\boldsymbol{B}$ matrices multiplied by the constants 0.75 and 0.80 for r equal to 3 and 5, respectively. The 'small' $\boldsymbol{B}$ matrix used in the simulations of Web Appendix Section 2.5 was calculated by multiplying the $r = 3$ large $\boldsymbol{B}$ by the constant 0.5. All other $p - 10$ rows of these $\boldsymbol{B}$ matrices were set to 0, where p is the total number of predictors used in the simulations. The $\boldsymbol{B}$ matrix used in Web Appendix Section 2.7 used the first 6 rows of the $r = 3$ 'moderate' $\boldsymbol{B}$ matrix, with all other $p - 5$ rows set to 0.

$$\boldsymbol{B}_{large,r=3}^{T} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ -2 & 2 & -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \end{bmatrix} \tag{3}$$

$$\boldsymbol{B}^T_{large,r=5} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ -2 & 2 & -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \\ -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ -1 & -1 & 0 & 1 & 1 & -1 & -1 & 0 & 1 & 1 & -2 \end{bmatrix} \tag{4}$$

The transpose of the first 6 rows of the deterministic 'moderate' $\boldsymbol{B}$ matrix used in the Poisson simulations in Web Appendix Section 2.8 are given in Web Appendix equation (5). All other $p - 5$ rows of the $\boldsymbol{B}$ matrices were set to 0, where $p = 100$ in the Poisson simulations.

$$\boldsymbol{B}^T_{poisson,r=3} = 0.75 \times \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 \end{bmatrix} \tag{5}$$

2.2 *Comparison of **glmmPen** vs **glmmPen_FA** methods*

As far as we are aware, the **glmmPen** method developed by Rashid et al. (2020) and implemented in the **glmmPen** R package available on CRAN is the only other method that performs simultaneous fixed and random effects variable selection in high dimensional GLMMs.

We next compare the performance of this **glmmPen** method and our novel **glmmPen_FA** method developed in this paper. We first compared the performance of these methods in moderate dimensions. We simulated binary responses from a logistic mixed effects model much like the procedure described in Section 3 of the main manuscript, except the total number of predictors used in the analyses was $p = 25$ and we restricted our consideration to $r = 3$ common factors for all simulation scenarios. For the **glmmPen_FA** method, all values of r used in the algorithm were from the Growth Ratio estimates of r. In these moderate dimensions of $p = 25$ with our given sample size of $N = 2500$, it is reasonable to use **glmmPen** to perform variable selection in logistic mixed effects models assuming an unstructured random effects covariance matrix, allowing us to use as directly comparable model assumptions as possible for these method comparisons.

Table 1 gives the average true and false positives for both the fixed and random effects, the median time in hours to complete the variable selection procedure, and the average of the mean absolute deviation between the coefficient estimates and the true coefficients across all simulation replicates. Table 2 gives the Growth Ratio r estimation procedure results for the **glmmPen_FA** method.

When comparing the **glmmPen_FA** and **glmmPen** results in these $p = 25$ simulations, we see that the median time for **glmmPen** to complete the variable selection procedures ranged from 2.53 to 3.28 hours for all four simulation scenarios considered. On the other hand, the **glmmPen_FA** method was able to fit these variable selection procedures about 5-6 times faster, where the median running time ranged from 0.47 to 0.59 hours.

Table 1 also shows that there is little difference in the true positives for both the fixed and random effects between the two methods. However, **glmmPen** tends to have more false positives in the fixed effects.

We also performed variable selection using **glmmPen** on the $r = 3$, $p = 100$ simulations described in Section 3 of the main manuscript. In these larger dimensions, we simplified the **glmmPen** estimation procedure by assuming an independent covariance matrix to reduce

β	$\boldsymbol{B}$	Method	TP % Fixef	FP % Fixef	TP % Ranef	FP % Ranef	T^{med}	Abs. Dev. (Mean)
1	Mod.	glmmPen_FA	99.50	3.47	99.80	0.53	0.59	0.26
		glmmPen	100.00	37.40	100.00	0.00	2.62	0.27
	Large	glmmPen_FA	97.20	3.80	99.90	0.80	0.49	0.33
		glmmPen	99.00	60.60	100.00	0.00	3.18	0.34
2	Mod.	glmmPen_FA	100.00	2.27	98.40	0.47	0.47	0.27
		glmmPen	100.00	13.67	99.20	0.00	2.84	0.43
	Large	glmmPen_FA	99.80	3.53	99.80	0.73	0.48	0.35
		glmmPen	100.00	30.93	100.00	0.00	2.53	0.50

Web Table 1: Results of the variable selection procedure for the $p = 25$ logistic mixed effects simulations comparing **glmmPen** with **glmmPen_FA**, 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 fixed effect coefficient estimates $\hat{\beta}$ 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 (moderate vs large). All values of r used in the glmmPen_FA method were from the Growth Ratio estimates of r.

True r	β	$\boldsymbol{B}$	Avg. r	r Underestimated %	r Correct %	r Overestimated %
3	1	Mod.	3.00	0	100	0
		Large	3.00	0	100	0
	2	Mod.	2.76	24	76	0
		Large	2.92	8	92	0

Web Table 2: Results of the Growth Ratio r estimation procedure for glmmPen_FA $p = 25$ logistic mixed effects simulations, including the average estimate of r across simulations and percent of times that the estimation procedure underestimated r, gave the true r, or overestimated r. 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 (moderate vs large).

the number of random effects covariance parameters. We let the **glmmPen** variable selection procedure run for 100 hours. In that time, **glmmPen** was able to complete the following number of replicates out of the 100 total replicates: 83 for ($\beta = 1, \boldsymbol{B} =$ Moderate), 71 for ($\beta = 1, \boldsymbol{B} =$ Large), 100 for ($\beta = 2, \boldsymbol{B} =$ Moderate), and 96 for for ($\beta = 2, \boldsymbol{B} =$ Large). The minimum times needed to complete the **glmmPen** variable selection procedures were 39.91, 57.60, 23.63, and 42.79 hours, respectively; in summary, it took a day or more for the fastest simulation replicates to complete when using the **glmmPen** method. In cases where we desire to select true random effects from a large number of total predictors, it is clear that the **glmmPen_FA** estimation procedure significantly reduces the required time to perform variable selection.

2.3 *Variable selection in Binomial data with 500 predictors*

In order to further illustrate the scalability of our method, we applied our method to binary outcome simulations with $p = 500$ covariates. We simulated the binary responses from a logistic mixed effects model much like the procedure described in Section 3 of the main manuscript, except the total number of predictors used in the analyses was $p = 500$ instead of $p = 100$. All simulations assumed the true number of latent factors r was 3 and the Growth Ratio method was used to estimate r. Just as in the $p = 100$ binary outcome simulations, we specified a full model for the algorithm such the candidate random effect predictors equalled the candidate fixed effect predictors (e.g. $q = p$), and our aim was to select the set of true predictors and random effects. The variable selection results to these simulations are given in Table 3. The median times needed to complete these simulations took between 10.37 and 17.57 hours.

True r	β	$\boldsymbol{B}$	Avg. r	TP % Fixef	FP % Fixef	TP % Ranef	FP % Ranef	T^{med}	Abs. Dev. (Mean)	$\|\|\boldsymbol{D}\|\|_F$
3	1	Mod.	2.67	97.40	0.97	93.90	0.05	10.37	0.27	1.15
		Large	2.85	88.50	1.73	94.30	0.33	17.57	0.37	2.52
	2	Mod.	2.37	100.00	0.06	77.40	0.00	11.44	0.48	0.69
		Large	2.41	99.60	0.19	88.10	0.03	13.22	0.55	1.35

Web Table 3: Results of the variable selection procedure for $p = 500$ logistic mixed effects simulations, 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 fixed effect coefficient estimates $\hat{\beta}$ and the true β values across all simulation replicates. Column 'r Avg.' gives the average Growth Ratio r estimate used within the algorithm. 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 (moderate vs large). 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.

2.4 *Variable selection simulations using the Elastic Net penalty applied to correlated predictors*

We extended our simulations on variable selection in Binomial data by adding correlations between the simulated covariates and adjusting for this correlation using the Elastic Net penalization approach. Elastic Net penalization balances the MCP, SCAD, or LASSO penalties with ridge regression. This balance between ridge regression and the other penalty is dictated by a value we label as π, where $\pi = 1.0$ represents the MCP penalty and $\pi = 0$ represents ridge regression.

In these simulations, we set the sample size to $N = 2500$ and number of groups to $K = 25$, with an equal number of subjects per group. There were $p = 100$ total predictors and 10 true predictors with non-zero fixed and random effects. We considered four types of correlations between the predictors. In three of the four correlation types, the correlation between all $p = 100$ covariates was set to a common value of 0.2, 0.4, or 0.6, and the variance of the covariates was set to 1.0. In the fourth correlation type, we randomly selected 100 of the 117 covariates used in the case study (see Web Appendix Section 3.1 for details) and calculated the Spearman correlation of these 100 covariates. In all four correlation cases, we simulated the covariates from a multivariate normal distribution with mean 0 and covariance matrix set to the correlation matrices described above.

We simulated the random effects covariance matrix using $r = 3$ and the corresponding moderate $\boldsymbol{B}$ matrix described in the Web Appendix Section 2.1. The 10 true fixed effects $\boldsymbol{\beta}$ coefficients were set to 1. The generation of the binary responses from a logistic mixed effects model proceeded as described in Section 3 of the main manuscript.

We performed variable selection on these simulated data using Elastic Net π values of 0.1, 0.3, 0.5, 0.8, and 1.0, and we estimated the number of common factors r using the default Growth Ratio procedure described in Section 2.4 of the main manuscript.

A summary of the variable selection results—true positive percentages, false positive percentages, median time in hours to complete the procedure, average absolute deviation between the estimated fixed effects coefficients and the true coefficients, and the average of the Frobenius norm of the difference between the estimated random effect covariance matrix $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^{\boldsymbol{T}}$ and the true covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^{\boldsymbol{T}}$ (the Frobenius norm was standardized by the number of random effects selected in the best model)—is given in Web Appendix Table 4. A summary of the performance of the Growth Ratio estimation procedure is given in Web Appendix Table 5.

In general, increasing the correlation among the predictors decreases the average true positive percentage. Within a particular correlation set-up, decreasing the value of the Elastic Net π tends to increase both the true positives and the false positives.

The Growth Ratio procedure tends to underestimate the number of common factors r as the correlation between the covariates increases. However, when the correlation between the covariates is high at a value of 0.6 and there is no adjustment for ridge regression (i.e. $\pi = 1.0$, equivalent to the MCP penalty), there are more instances of the Growth Ratio procedure overestimating r.

For low values of π and/or high correlation, some simulation replicates had model fit issues. Specifically, in certain situations, the random effect variances diverged to excessively large values. As a result, the BIC-ICQ model selection criteria could not be calculated for the model, and the model selection procedure was suspended. When $\pi = 0.1$ and the correlation among the predictors was 0.4 or 0.6, this phenomena happened 25% or 26% of the time, respectively. When $\pi = 0.1$ and the correlation was 0.2, this happened 2% of the time; when $\pi = 0.3$ and the correlation was 0.4 or 0.6, this happened 1% or 3% of the time, respectively; when $\pi = 1.0$ and the correlation was 0.6, this happened 1% of the time. The simulations summarized in Web Appendix Table 4 do not include results from these problematic simulation replicates.

2.5 *Variable selection simulations with alternative β and B matrix sizes*

We ran additional $p = 100$ logistic mixed effects variable selection simulations with alternative combinations of predictor effects ($\beta = 0.5, 1.0, 2.0$) and $\boldsymbol{B}$ matrix size (small, see Web Appendix Section 2.1). We simulated the binary responses from a logistic mixed effects model much like the procedure described in Section 3 of the main paper, except we modified the $\boldsymbol{\beta}$ and $\boldsymbol{B}$ combinations considered. All simulations assumed the true number of latent factors r was 3 and the Growth Ratio method was used to estimate r.

Corr	π	TP % Fixef	FP % Fixef	TP % Ranef	FP % Ranef	T^{med}	Abs. Dev. (Mean)	$\|\|\boldsymbol{D}\|\|_F$
0.2	0.1	97.55	27.39	85.10	12.40	23.34	0.50	0.98
	0.3	91.70	13.32	61.90	4.46	22.06	0.54	1.29
	0.5	93.30	7.79	65.80	2.64	8.63	0.51	1.26
	0.8	94.10	1.08	90.00	0.63	2.58	0.40	1.71
	1.0	94.70	4.28	94.40	0.53	1.46	0.31	1.31
0.4	0.1	86.53	22.79	74.53	14.33	14.37	0.56	2.69
	0.3	87.98	11.93	59.49	1.99	15.10	0.58	1.21
	0.5	85.70	4.64	72.70	1.07	4.71	0.53	1.08
	0.8	80.30	1.61	83.30	0.37	2.54	0.44	0.74
	1.0	80.40	1.98	82.60	0.21	1.17	0.39	0.90
0.6	0.1	80.14	17.16	53.24	7.03	13.08	0.63	2.76
	0.3	76.39	10.63	55.26	3.13	10.31	0.61	1.22
	0.5	75.00	4.20	58.50	0.86	4.10	0.55	1.12
	0.8	76.70	1.28	64.20	0.61	2.47	0.45	0.79
	1.0	71.31	0.85	56.26	0.24	1.40	0.41	0.74
CS	0.1	93.10	45.41	92.00	29.90	28.90	0.49	1.01
	0.3	93.80	28.63	74.30	16.32	24.74	0.46	1.16
	0.5	87.90	19.19	66.60	11.71	12.78	0.44	1.33
	0.8	91.70	3.99	75.00	2.94	3.12	0.40	1.32
	1.0	95.90	2.04	86.20	1.38	1.36	0.30	2.72

Web Table 4: Variable selection results for the Elastic Net $p = 100$ logistic mixed effects simulations, 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 fixed effect coefficient estimates $\hat{\beta}$ and the true β values across all simulation replicates. Column "Corr" describes the correlation between the covariates (equal correlation of values 0.2, 0.4, or 0.6, or correlation based on data from the case study data, labeled as 'CS'). Column π represents the Elastic Net balance between ridge regression ($\pi = 0$) and the MCP penalty ($\pi = 1$). 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.

We ran these simulations to demonstrate how our method performed under the following scenarios:

(1) Fixed effects coefficients remained the same, but size of the random effect variation decreased
(2) Random effect variation decreased AND the fixed effects coefficients decreased

The variable selection results for these simulations are given in Web Appendix Table 6. From the results presented in Web Appendix Table 6, we see that decreasing the value of the fixed effects coefficients β while keeping the size of the $\boldsymbol{B}$ matrix consistently 'small' resulted in decreased fixed effect true positive rates and increased random effect true positive rates.

Corr	π	Avg. r	r Underestimated %	r Correct %	r Overestimated %
0.2	0.1	2.67	35	63	2
	0.3	2.65	35	65	0
	0.5	2.63	37	63	0
	0.8	2.54	46	54	0
	1.0	2.62	39	60	1
0.4	0.1	2.24	76	24	0
	0.3	2.35	67	32	1
	0.5	2.38	68	28	4
	0.8	2.47	69	26	5
	1.0	2.35	73	24	3
0.6	0.1	2.05	95	5	0
	0.3	2.30	82	12	5
	0.5	2.30	81	12	7
	0.8	2.38	84	8	8
	1.0	2.99	72	13	15
CS	0.1	2.47	53	47	0
	0.3	2.45	55	45	0
	0.5	2.42	58	42	0
	0.8	2.44	56	44	0
	1.0	2.43	57	43	0

Web Table 5: Results of the Growth Ratio r estimation procedure for the Elastic Net $p = 100$ logistic mixed effects simulations, including the average estimate of r across simulations and percent of times that the estimation procedure underestimated r, gave the true r, or overestimated r. Column "Corr" describes the correlation between the covariates (equal correlation of values 0.2, 0.4, or 0.6, or correlation based on data from the case study data). Column π represents the Elastic Net balance between ridge regression ($\pi = 0$) and the MCP penalty ($\pi = 1$).

This increase in the true positive rate for the random effects was likely a consequence of the increased accuracy of the estimation of r. This pattern is consistent with findings from Tables 1 and 2 of the main paper, which also show that decreasing the predictor effect β decreases the fixed effect true positive rates, increases the random effect true positive rates, and improves the accuracy in the estimation of r using the Growth Ratio procedure.

If we compare the results in Web Appendix Table 2.5 with Table 1 from the main paper, we see that decreasing the size of the $\boldsymbol{B}$ matrix decreased the true positive rates of the random effects, but otherwise had minimal impact on the fixed effect true positive rates.

2.6 *Variable selection simulations with alternative sample size and number of groups*

We ran additional $p = 100$ logistic mixed effects variable selection simulations with a smaller overall sample size of $N = 1000$ (compared with the $N = 2500$ for simulations in the main paper) and $K = 10, 25$ groups. We simulated the binary responses from a logistic mixed effects model much like the procedure described in Section 3 of the main paper, except we modified the overall sample size and the number of groups considered. All simulations assumed the true number of latent factors r was 3 and the Growth Ratio method was used to estimate r.

True r	β	$\boldsymbol{B}$	Avg. r	TP % Fixef	FP % Fixef	TP % Ranef	FP % Ranef	T^{med}	Abs. Dev. (Mean)	$\|\|\boldsymbol{D}\|\|_F$
3	0.5	Small	2.78	91.20	3.10	95.90	0.36	1.57	0.15	0.31
	1.0		2.42	100.00	1.11	90.50	0.37	1.52	0.19	0.34
	2.0		2.13	100.00	0.94	74.60	0.24	1.07	0.24	0.44

Web Table 6: Results of the variable selection procedure for $p = 100$ logistic mixed effects simulations using alternative $\boldsymbol{\beta}$ and $\boldsymbol{B}$ magnitudes. 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 fixed effect coefficient estimates $\hat{\beta}$ and the true β values across all simulation replicates. Column 'r Avg.' gives the average Growth Ratio r estimate used within the algorithm. 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 (small). 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.

We ran these simulations to demonstrate how our method performed under the following scenarios:

(1) Overall sample size decreases and the group sample sizes decrease (i.e. the number of groups remains the same)

(2) Overall sample size decreases and the group sample sizes remain the same (i.e. the number of groups decrease)

The variable selection results to these simulations are given in Web Appendix Table 7. When we compare the results presented in Web Appendix Table 7 with the results presented in Table 1 of the main paper, we see that decreasing the overall sample size from $N = 2500$ to $N = 1000$ while keeping $K = 25$ generally increased the false positive rates for both the fixed and random effects; generally decreased the true positive rate of the random effects; and generally increased the error in the random effect covariance matrix calculation (i.e. the Frobenius norm of the difference between the estimated and true random effects covariance matrix increased). When K was reduced to 10, the false positive rates for both the fixed and random effects increased further, as did the error in the random effect covariance matrix calculation. Overall, we can conclude that having a greater number of groups within the data and increasing the sample size per group both help improve the selection results.

2.7 *Variable selection simulations with alternative number of true random effects*

We ran additional $p = 100$ logistic mixed effects variable selection simulations with a smaller number of true random effects. Instead of setting all 10 true predictors to have both non-zero fixed and non-zero random effects, we set all 10 true predictors to have non-zero fixed effects but only 5 of the 10 true predictors to have non-zero random effects. We simulated the binary responses from a logistic mixed effects model much like the procedure described in Section 3 of the main paper, except we modified which random effect predictors were included in the

True r	$\boldsymbol{\beta}$	$\boldsymbol{B}$	K	Avg. r	TP % Fixef	FP % Fixef	TP % Ranef	FP % Ranef	T^{med}	Abs. Dev. (Mean)	$\|\|\boldsymbol{D}\|\|_F$
3	1	Mod.	10	2.47	91.30	15.22	73.00	9.72	1.48	1.33	11.47
			25	2.58	95.50	8.57	66.90	2.89	1.09	0.80	5.69
		Large	10	2.60	84.90	22.16	81.50	15.84	3.14	2.02	22.88
			25	2.50	89.80	18.84	81.00	9.72	1.98	1.43	17.44
	2	Mod.	10	2.24	99.80	12.12	69.30	5.36	1.83	1.87	7.15
			25	3.78	99.90	7.98	44.20	3.56	3.67	1.91	6.13
		Large	10	2.28	98.90	16.54	76.20	6.84	2.79	2.08	15.14
			25	2.92	99.90	15.21	70.40	6.77	2.74	1.85	10.04

Web Table 7: Results of the variable selection procedure for $p = 100$ logistic mixed effects simulations using alternative combinations of sample size $N = 1000$ and number of groups K. 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 fixed effect coefficient estimates $\hat{\beta}$ and the true β values across all simulation replicates. Column 'r Avg.' gives the average Growth Ratio r estimate used within the algorithm. 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 (moderate vs large). 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.

simulation of the outcome and we modified the $\boldsymbol{B}$ matrix used (see Web Appendix Section 2.1). All simulations assumed the true number of latent factors r was 3; the Growth Ratio method was used to estimate r; and the predictor effects were moderate ($\beta = 1$).

We ran this simulation to demonstrate how our method performed when the true number of random effects were smaller than the true number of fixed effects.

The variable selection results for this simulation are given in Web Appendix Table 8. When we compare these results presented in Web Appendix Table 8 with the comparable results presented in Table 1 of the main paper, we see that decreasing the number of true random effects in the model resulted in decreased true positive rates for the random effects; this was likely a consequence of the decreased accuracy of the estimation of r when the true number of random effects in the model decreased. The true and false positive rates of the fixed effects remained consistent with the main paper Table 1 results.

2.8 *Variable selection in Poisson data with 100 predictors*

While we have previously focused on binary outcome data in our simulations, our proposed method also applies to other members of the generalized linear model family, including the Poisson model for count outcome data. To illustrate this, we simulated a Poisson mixed effects model with $p = 100$ predictors, 5 of which had truly non-zero fixed and random effects, and the other $p - 5$ predictors had zero-valued fixed and random effects. As in the previous Binomial simulations, we set the sample size to $N = 2500$ and the number of groups to $K = 25$, with equal numbers of subjects per group. We set r to 3, assigned moderate predictor

True r	β	$\boldsymbol{B}$	Avg. r	TP % Fixef	FP % Fixef	TP % Ranef	FP % Ranef	T^{med}	Abs. Dev. (Mean)	$\|\|\boldsymbol{D}\|\|_F$
3	1	Mod.	2.00	99.60	1.93	78.80	1.28	1.17	0.17	0.77

Web Table 8: Results of the variable selection procedure for $p = 100$ logistic mixed effects simulations using an alternative numbers of true random effects (5 true random effects predictors, which were a subset of the 10 true fixed effects predictors). 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 fixed effect coefficient estimates $\hat{\beta}$ and the true β values across all simulation replicates. Column 'r Avg.' gives the average Growth Ratio r estimate used within the algorithm. 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 (moderate). 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.

effects ($\beta = 1$), and specified a $\boldsymbol{B}$ matrix with 6 non-zero rows (for the 5 predictors plus an intercept) that produced a 'moderate' covariance matrix (see Web Appendix Section 2.1 for details). Unlike in the previous Binomial simulations, we simulated $x_{ki,j} \sim N(0, \sigma = 0.10)$ for $j = 1, ..., p$ to reduce the overall spread of the simulated $y_{ki} \sim Poisson(\mu_{ki})$ outcome values, where $\mu_{ki} = \exp(\boldsymbol{x}_{ki}^T\boldsymbol{\beta} + \boldsymbol{z}_{ki}^T\boldsymbol{\gamma}_k)$. See Web Appendix Section 1.3 for details on the penalty sequences used.

Using the Growth Ratio estimation procedure to estimate r, the average true positive percentages were 83.40% for the fixed effects and for 77.20% the random effects, and the average false positives were 5.92% for the fixed effects and 3.23% for the random effects. The average estimate of r across the simulation replicates was 2.35.

3. Web Appendix: Supplementary case study details and materials

This section of the Web Appendix provides additional details about the case study data preparation procedure and the case study sensitivity analyses.

3.1 *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_glmmPen_FA`.

The studies used in these analyses are summarized in Web Appendix Table 9, which contains the gene expression platform information (all RNA-seq), the sample sizes, and the percent of the samples that were classified into the basal subgroup.

Web Appendix Table 9 provides the dataset abbreviations, their respective citations, their sample sizes, and the percent of the subjects within each study that were classified into the basal subtype. The sample sizes listed in the table—and used in the analyses—were

Dataset	Platform	Sample Size	% Basal	Citation
Aguirre	RNA-seq	28	29	Aguirre et al. (2018)
CPTAC	RNA-seq	99	22	Cao et al. (2021)
Dijk	RNA-seq	61	39	Dijk et al. (2020)
Hayashi	RNA-seq	75	53	Hayashi et al. (2020)
TCGA	RNA-seq	97	20	Raphael et al. (2017)

Web Table 9: Summaries of the five pancreatic ductal adenocarcinoma (PDAC) gene expression datasets used in the case study analyses described in Section 4 of the main paper.

smaller than the studies' total sample size as we removed subjects with missing tumor grade information, normal tissue samples, and those who did not have primary tumor samples of sufficient quality. All five datasets had RNA-seq data for 60,230 total gene symbols for each subject. Of those, 432 were also part of the 500 member gene list that Moffitt et al. (2015) identified as likely to be expressed solely in PDAC tumor cells (this gene list is also provided within the aforementioned GitHub repository).

There were some significant correlations between some of these 432 genes, as evaluated by Spearman correlations applied to the subjects' rank transformed gene expression. In order to avoid having very highly correlated covariates in the analyses, we decided to combine highly correlated genes together into meta-genes. We accomplished this by applying a hierarchical clustering algorithm—the `pheatmap::pheatmap()` R function (Kolde, 2019)—to the Spearman correlation matrix of the rank-transformed gene expression. We then cut the tree using the `stats::cutree()` R function (R Core Team, 2021) at a height of 3. This produced a total of 117 clusters, or meta-genes. For meta-genes that represented two or more genes, we added the raw RNA-seq gene expression of all participating genes, which were then rank-transformed on the subject level; these rank-transformed meta-gene covariates were used in the case study analyses.

The cancer subtype outcome—basal or classical—was calculated using the clustering algorithm specified in Moffitt et al. (2015). For each study individually, this clustering algorithm was applied to the RNA-seq gene expression for the 432 genes described above, where the distance matrix was the Euclidean distance and the assumed number of clusters was set to two.

3.2 *Case study: Sensitivity analyses*

Using both **glmmPen_FA** and **glmmPen**, 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 **glmmPen_FA** procedure). Based on the results in Web Appendix Table 4, π values between 0.5 and 1.0 were likely to have good selection results for the correlation structure of the covariates in the dataset. We fit the variable selection procedure using $\pi = \{0.6, 0.7, 0.8, 0.9, 1.0\}$. The **glmmPen** procedure assumed an independent random effects covariance matrix.

We first discuss the **glmmPen_FA** sensitivity results. In addition to estimating the number of latent common factors using the Growth Ratio procedure, which estimated a value of $r = 2$ for all values of π, we also fit the model assuming $r = 3$ because the simulations given in

Web Appendix Section 2.4 indicated that the Growth Ratio method may underestimate r. Regardless of whether r was estimated as 2 using the Growth Ratio procedure or set to 3 manually, the coefficient values and selection results were very consistent for each value of π. The single exception was when the $\pi = 0.7$ selection procedure included two additional meta-gene covariate in the best model for $r = 2$ but not $r = 3$ (meta-gene 36, gene KLF5, positive log odds ratio; meta-gene 111, genes DMKN, RHOV, VGLL1, positive log odds ratio). Because of these similarities, we restricted our consideration to models where r was set to the Growth Ratio estimate of 2.

In terms of fixed effects, the values π between 0.6 and 1.0 gave very consistent results within the **glmmPen_FA** procedure ($r = 2$). The 8 covariate meta-genes described in the main paper Table 3 (which describe the **glmmPen_FA** results using $\pi = 0.8$ and $r = 2$ estimated from the Growth Ratio procedure) were consistently chosen across the different values of π, with the exception of meta-gene 7, which was excluded from the best model when $\pi = \{1.0\}$. The meta-genes 36 and 111 (described above) were also selected when $\pi = 0.7$; meta-gene 36 was also selected when $\pi = 0.6$; and meta-gene 111 was also selected when $\pi = 1.0$.

For random effects, all values of π (and all values of r) consistently selected 0 random effect slopes (random intercept only for random effects) within the **glmmPen_FA** procedure. The random intercept variances ranged between 0.27 and 0.84 when $r = 2$, with variances increasing as the value of π increased.

We chose to report the **glmmPen_FA** results of $\pi = 0.8$ in the main manuscript for several reasons. Based on the fact that we had a range of correlations among the covariates in the dataset, including some pairwise correlations greater than 0.5, we felt it was appropriate to fit the variable selection procedure with $\pi < 1.0$. When choosing between the other values of π, the $\pi = 0.8$ results contained the consistently selected 8 meta-gene covariates.

The times to complete the **glmmPen_FA** variable selection procedure was between 0.4 and 1.8 hours for $\pi = \{0.6, 0.7, 0.8, 0.9, 1.0\}$ (this range includes r values equal to 2 or 3).

The **glmmPen** procedure also consistently selected the 8 covariate meta-genes described in the main paper Table 3 across all values of $\pi = \{0.6, 0.7, 0.8, 0.9\}$; the $\pi = 1.0$ fit did not complete within 4 days (96 hours) and is therefore not included in these sensitivity analyses. When $\pi = \{0.6, 0.7, 0.8\}$, **glmmPen** consistently selected the additional meta-genes 59 (genes PKIB, DNAJC15, negative log odds ratio) and 71 (genes AKR1C3, CA2, MGST2; positive log odds ratio). The **glmmPen** method consistently selected meta-gene 117 to have a non-zero random effect (variance ranged between 0.46 and 1.11, with variances increasing as the value of π increased). For $\pi = 0.9$, **glmmPen** also selected meta-gene 7 to have a non-zero random effect (variance 0.69).

The times in hours to complete the **glmmPen** variable selection procedure was 53.5, 57.7, 49.2, and 60.2 for π equal to 0.6, 0.7, 0.8, and 0.9, respectively.

References

Aguirre, A. J., Nowak, J. A., Camarda, N. D., Moffitt, R. A., Ghazani, A. A., Hazar-Rethinam, M., Raghavan, S., Kim, J., Brais, L. K., Ragon, D., et al. (2018). Real-time genomic characterization of advanced pancreatic cancer to enable precision medicine. *Cancer discovery* **8,** 1096–1111.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics* **5,** 232–253.

Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and computing* **25,** 173–187.

Cao, L., Huang, C., Zhou, D. C., Hu, Y., Lih, T. M., Savage, S. R., Krug, K., Clark, D. J., Schnaubelt, M., Chen, L., et al. (2021). Proteogenomic characterization of pancreatic ductal adenocarcinoma. *Cell* **184,** 5031–5052.

Dijk, F., Veenstra, V. L., Soer, E. C., Dings, M. P., Zhao, L., Halfwerk, J. B., Hooijer, G. K., Damhofer, H., Marzano, M., Steins, A., et al. (2020). Unsupervised class discovery in pancreatic ductal adenocarcinoma reveals cell-intrinsic mesenchymal features and high concordance between existing classification systems. *Scientific reports* **10,** 337.

Hayashi, A., Fan, J., Chen, R., Ho, Y.-j., Makohon-Moore, A. P., Lecomte, N., Zhong, Y., Hong, J., Huang, J., Sakamoto, H., et al. (2020). A unifying paradigm for transcriptional heterogeneity and squamous features in pancreatic ductal adenocarcinoma. *Nature Cancer* **1,** 59–74.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67,** 495–503.

Kolde, R. (2019). *pheatmap: Pretty Heatmaps.* R package version 1.0.12.

Moffitt, R. A., Marayati, R., Flate, E. L., Volmar, K. E., Loeza, S. G. H., Hoadley, K. A., Rashid, N. U., Williams, L. A., Eaton, S. C., Chung, A. H., et al. (2015). Virtual microdissection identifies distinct tumor-and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature genetics* **47,** 1168.

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Raphael, B. J., Hruban, R. H., Aguirre, A. J., Moffitt, R. A., Yeh, J. J., Stewart, C., Robertson, A. G., Cherniack, A. D., Gupta, M., Getz, G., et al. (2017). Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer cell* **32,** 185–203.

Rashid, N. U., Li, Q., Yeh, J. J., and Ibrahim, J. G. (2020). Modeling between-study heterogeneity for improved replicability in gene signature selection and clinical prediction. *Journal of the American Statistical Association* **115,** 1125–1138.