# Supporting Information for "Dimension reduction for integrative survival analysis"

Aaron J. Molstad*† and Rohit K. Patra*
Department of Statistics* and Genetics Institute†, University of Florida
Gainesville, Florida, U.S.A.
amolstad@ufl.edu and rohitpatra@ufl.edu

# A Convex approximation

## A.1 Proposed estimator

As both $\mathcal{C}_r$ and $\mathcal{A}_s$ are nonconvex sets, the optimization problem in (3) is nonconvex. Hence, there is no guarantee that our algorithm converges to a global minimizer. A common alternative is to approximate (nonconvex) $L_0$ constraints with (convex) $L_1$ constraints. Applied to (3), this would correspond to the estimator

$$\underset{\boldsymbol{B}\in\mathbb{R}^{p\times J}}{\arg\min}\left\{-\boldsymbol{\mathcal{L}}(\boldsymbol{B})+\gamma_1\|\boldsymbol{B}\|_*+\gamma_2\|\boldsymbol{B}\|_{1,2}\right\} \tag{11}$$

where $\|\cdot\|_*$ is the nuclear norm (i.e., the norm which sums the singular values of its matrix-valued argument), $\|\boldsymbol{A}\|_{1,2}=\sum_{j=1}^{p}\|\boldsymbol{A}_{j,\cdot}\|_2$, and $(\gamma_1,\gamma_2)\in[0,\infty)\times[0,\infty)$ are tuning parameters. Large values of $\gamma_1$ will force many singular values of $\boldsymbol{B}$ to be zero, thus reducing its rank. Similarly, when $\gamma_2$ is large, the second penalty will force many entire rows of $\boldsymbol{B}$ to be zero. The estimator (11) can thus be thought of as a convex approximation to (3).

While this estimator may seem appealing, we encounter three issues with (11). First, the use of two penalties often leads to over-shrinkage towards the origin. Second, identifying a tuning parameter pair that leads to both low-rank and sparse estimates of $\boldsymbol{B}_*$ was difficult in the settings we considered. To choose a $\gamma_1$ which led to low-rankness, it was most often the case that (11) needed to be entirely nonzero, whereas to achieve a sparse estimate, (11) often needed to be of full rank. Third, computing (11) requires solving a difficult optimization problem. To the best of our knowledge, there exists no standard algorithm for solving optimization problems like (11). In the following subsection, we propose a proximal-proximal gradient descent algorithm (Ryu and Yin, 2019) for computing (11) that may be of independent interest. In our simulation studies, we show that despite the virtues of convexity, (3) outperforms (11) to a notable extent.

## A.2 Computing the convex approximation

As mentioned in preceding section, the optimization problem in (11) is convex. Solving this optimization problem is especially challenging because the penalties are a function of all columns of $\boldsymbol{B}$, making the optimization problem nonseparable across populations. However, like many penalized maximum likelihood estimators, the corresponding optimization problem can be expressed as the sum of differentiable and nondifferentiable functions, which suggests that we may employ "splitting algorithms" to solve (11). We use the so-called proximal-proximal gradient descent algorithm (Ryu and Yin, 2019), also known as the Davis-Yin splitting algorithm (Davis and Yin, 2017), for computing (11) with $\gamma_1 > 0$ and $\gamma_1 > 0$.

Throughout this section, we refer to the proximal operator of a function $h$, denoted $\mathbf{prox}_h$, which we define as

$$\mathbf{prox}_h(\boldsymbol{B}) = \arg\min_{\boldsymbol{A}} \left\{ \frac{1}{2}\|\boldsymbol{A} - \boldsymbol{B}\|_F^2 + h(\boldsymbol{A}) \right\}.$$

To employ the proximal-proximal gradient descent algorithm, we need only be able to evaluate the proximal operators of $\|\cdot\|_*$ and $\|\cdot\|_{1,2}$ separately, and compute the gradient of $\boldsymbol{\mathcal{L}}$ with respect to $\boldsymbol{B}$. Specifically, we update three sets of variables at each iteration: $\boldsymbol{B} \in \mathbb{R}^{p\times J}$, $\widetilde{\boldsymbol{B}} \in \mathbb{R}^{p\times J}$, and $\boldsymbol{\Theta} \in \mathbb{R}^{p\times J}$. Given $\alpha > 0$, a sufficiently small step size parameter, the updating equations to obtain the $(t)$th set of iterates are

$$\begin{aligned}
\boldsymbol{B}^{(t)} &= \mathbf{prox}_{\alpha\gamma_1\|\cdot\|_*}(\boldsymbol{\Theta}^{(t-1)}), \\
\widetilde{\boldsymbol{B}}^{(t)} &= \mathbf{prox}_{\alpha\gamma_2\|\cdot\|_{1,2}}\{2\boldsymbol{B}^{(t)} - \boldsymbol{\Theta}^{(t-1)} + \alpha\nabla\mathcal{L}(\boldsymbol{B}^{(t)})\}, \\
\boldsymbol{\Theta}^{(t)} &= \boldsymbol{\Theta}^{(t-1)} + \widetilde{\boldsymbol{B}}^{(t)} - \boldsymbol{B}^{(t)}.
\end{aligned}$$

To lend intuition to this procedure, note that if $\gamma_1 = 0$, then $\boldsymbol{B}^{(t)} = \boldsymbol{\Theta}^{(t-1)}$, so that iterates above would be exactly the iterates of the standard proximal gradient descent algorithm with step size parameter $\alpha$ (Parikh and Boyd, 2014, Section 4.2). However, applying proximal gradient descent to (11) with both $\gamma_1 > 0$ and $\gamma_2 > 0$ directly would be problematic because this would require computing the proximal operator of the sum of two penalties, which would require its own iterative procedure.

The proximal-proximal gradient descent algorithm above is particularly efficient because each of the iterates can be computed in closed form. In particular, for a matrix $\boldsymbol{A}$,

$$\mathbf{prox}_{\tau\|\cdot\|_*}(\boldsymbol{A}) = \boldsymbol{\Lambda}\mathbf{soft}(\boldsymbol{\Phi}, \tau)\boldsymbol{\Gamma}^\top, \tag{12}$$

where $\boldsymbol{A} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Gamma}^\top$ is the singular value decomposition of $\boldsymbol{A}$ (where $\boldsymbol{\Phi}$ is diagonal with nonnegative entries, $\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda} = I$, and $\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma} = I$) and **soft** is the soft thresholding operator applied elementwise to its argument, i.e., for all pairs $(j, k)$,

$$[\mathbf{soft}(\boldsymbol{C}, \tau)]_{j,k} = \max\left(|\boldsymbol{C}_{j,k}| - \tau, 0\right)\operatorname{sign}(\boldsymbol{C}_{j,k}).$$

For a derivation of (12), see the proof of Proposition 1 of Zhou and Li (2014), for example. Similarly, the proximal operator of the $\|\cdot\|_{1,2}$ norm also has a closed form. Letting

$[\mathbf{prox}_{\tau\|\cdot\|_{1,2}}(\boldsymbol{A})]_{j,\cdot}$ denote the $j$th row of the proximal operator of $\tau\|\cdot\|_{1,2}$ evaluated at $\boldsymbol{A} \in \mathbb{R}^{a\times b}$, we have

$$[\mathbf{prox}_{\tau\|\cdot\|_{1,2}}(\boldsymbol{A})]_{j,\cdot} = \begin{cases} \boldsymbol{A}_{j,\cdot}\left(1 - \frac{\tau}{\|\boldsymbol{A}_{j,\cdot}\|_2}\right) & : \|\boldsymbol{A}_{j,\cdot}\|_2 > \tau \\ 0 & : \|\boldsymbol{A}_{j,\cdot}\|_2 \leq \tau \end{cases}, \quad j \in [a].$$

Applying a result from Davis and Yin (2017), it follows that if $\alpha > 0$ is fixed sufficiently close to zero – according to the Lipschitz constant of $\nabla\boldsymbol{\mathcal{L}}$ – then the sequence of iterates defined above satisfy $\boldsymbol{B}^{(t)} \to \boldsymbol{B}^\star$, $\widetilde{\boldsymbol{B}}^{(t)} \to \boldsymbol{B}^\star$ as $t \to \infty$ where $\boldsymbol{B}^\star$ is a solution to (11).

While the proximal-proximal gradient descent algorithm is easy to implement, fixing the step size $\alpha$ can often lead to slow convergence in practice. In our implementation, we use a version of this algorithm proposed by Pedregosa and Gidel (2018) which selects step sizes adaptively using a backtracking line search. We found that in practice, this approach led to much shorter computing times than the version with $\alpha$ fixed. Specifically, we implement a version of Variant 1 of Algorithm 1 from Pedregosa and Gidel (2018) applied to (11).

When either $\gamma_1 = 0$ or $\gamma_2 = 0$, we instead use an accelerated proximal gradient descent algorithm (Parikh and Boyd (2014), Sections 4.2–4.3) to solve (11). For more on the accelerated proximal gradient descent algorithm, see Beck and Teboulle (2009).

# B Tuning parameter selection

There are numerous criteria used to selecting tuning parameters for fitting penalized Cox proportional hazards models. For example, `glmnet` uses cross-validated partial likelihood. In our implementation, however, we instead implement a cross-validated linear predictors criterion proposed in Dai and Breheny (2019). Specifically, for each $j \in [J]$, we randomly assign subjects to one of $V$ folds, $\mathcal{K}_{(j)1}, \ldots, \mathcal{K}_{(j)V}$, where $\mathcal{K}_{(j)1}, \ldots, \mathcal{K}_{(j)V}$ is a partition of $[n_{(j)}]$. For the $v$th fold, we compute $\widehat{\boldsymbol{B}}_{\widetilde{r},\widetilde{s},(v)}$ by fitting (3) to all subjects not belonging to the $v$th fold with $s = \widetilde{s}$ and $r = \widetilde{r}$. Then, for each $i \in \mathcal{K}_{(j)v}$, we compute and store

$$\phi_{(j)i,(\widetilde{r},\widetilde{s})} = \boldsymbol{x}_{(j)i}^\top \widehat{\boldsymbol{B}}_{\widetilde{r},\widetilde{s},(v)} \quad j \in [J],$$

for each $v \in [V]$. Thus, with $\phi_{(j)1,(\widetilde{r},\widetilde{s})}, \ldots, \phi_{(j)n_{(j)},(\widetilde{r},\widetilde{s})}$ in hand, we select the tuning pair $(\widehat{r}, \widehat{s})$ which maximizes the (weighted) partial log-likelihood

$$(\widehat{r}, \widehat{s}) = \underset{(\widetilde{r},\widetilde{s})\in\mathcal{T}}{\arg\max} \frac{1}{J}\sum_{j=1}^{J}\left[w_{(j)}\sum_{i=1}^{n_{(j)}}\delta_{(j)i}\log\left\{\frac{\exp(\phi_{(j)i,(\widetilde{r},\widetilde{s})})}{\sum_{k\in\mathcal{R}_{(j)i}}\exp(\phi_{(j)k,(\widetilde{r},\widetilde{s})})}\right\}\right], \tag{13}$$

where $\mathcal{T}$ is a prespecified (discrete) set of candidate tuning parameters and $w_{(j)} \geq 0$ are weights corresponding to the $j$th dataset. In our simulations and real data analysis, we set $w_{(j)} = 1$, but there may be settings where one would prefer to use $w_{(j)} \neq n_{(j')}$ for some pairs $j \neq j'$. We found (13) worked better than cross-validated partial likelihood, particularly when many subjects' survival times were censored. This is partly because cross-validated linear predictors do not require constructing a risk set separately for each fold: see Dai and Breheny (2019) for more on this approach.

Note also that one should be mindful of the tuning parameter $\mu$ and initial penalty parameter $\rho_0$ when performing $V$-fold cross-validation. When leaving out the $v$th fold for model fitting, the effect of $\mu$ and $\rho$ will be different because the magnitude of the partial log-likelihood depends on the sample size. Roughly speaking, $\mu$ and $\rho_0$ should be scaled by $(V-1)/V$ during $V$-fold cross-validation, and should be returned to their original values for fitting the model to the entire dataset. For example, in Section 6, we set $\mu \approx (4/5)50$ during 5-fold cross-validation, and set $\mu = 50$ for fitting the model to the complete dataset.

# C Simulation study performance metrics

We use three performance metrics in our simulation studies: model error, C-index, and Brier score.

Model error, defined as $\mathrm{tr}\{(\widehat{\boldsymbol{B}} - \boldsymbol{B}_*)^\top \boldsymbol{\Sigma} (\widehat{\boldsymbol{B}} - \boldsymbol{B}_*)\}$, quantifies the accuracy of the linear predictors. In the context of our simulation study, this can be thought of as $\sum_{j=1}^{J} \lim_{n_{(j)} \to \infty} \|\boldsymbol{X}_{(j)}(\boldsymbol{b}_{*(j)} - \widehat{\boldsymbol{b}}_{(j)})\|_2^2 / n_{(j)}$. C-index, in contrast, measures the degree of agreement in ordering between the linear predictor and the observed event times. As no outcomes are censored in our simulation study testing set, the C-index between the linear predictors and the observed outcomes can simply be expressed

$$\frac{1}{J} \sum_{j=1}^{J} \sum_{i=1}^{n_{\text{test}}} \sum_{k=1}^{n_{\text{test}}} \frac{\mathbf{1}(y_{(j)i} > y_{(j)k}) \{ \mathbf{1}(\boldsymbol{x}_{(j)i}^\top \widehat{\boldsymbol{b}}_{(j)} < \boldsymbol{x}_{(j)k}^\top \widehat{\boldsymbol{b}}_{(j)}) + \frac{1}{2} \times \mathbf{1}(\boldsymbol{x}_{(j)i}^\top \widehat{\boldsymbol{b}}_{(j)} = \boldsymbol{x}_{(j)k}^\top \widehat{\boldsymbol{b}}_{(j)}) \}}{\sum_{s=1}^{n_{\text{test}}} \sum_{t=1}^{n_{\text{test}}} \mathbf{1}(y_{(j)s} > y_{(j)t})}.$$

Note that a C-index of one indicates perfect agreement in ordering between observed outcomes and estimated linear predictors, whereas a C-index of 0.5 suggests that a model is no better than randomly guessing the order of the linear predictors. Hence, unlike model error, a higher C-index indicates better performance.

Finally, we also measure the Brier score evaluated at the median observed survival time. When there is no censoring, the Brier score for the $j$th population at time $t$ is defined as

$$B(t \mid \widehat{\boldsymbol{b}}_{(j)}) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left\{ \mathbf{1}(y_{(j)i} > t) - \widehat{S}_{(j)}(t \mid \boldsymbol{x}_{(j)i}, \widehat{\boldsymbol{b}}_{(j)}) \right\}^2,$$

where $\widehat{S}_{(j)}(t \mid \boldsymbol{x}_{(j)i}, \widehat{\boldsymbol{b}}_{(j)})$ is the estimated survival probability at time $t$ in the $j$th population based on estimate $\widehat{\boldsymbol{b}}_{(j)}$ for a subject with predictors $\boldsymbol{x}_{(j)i}$. Then, the (averaged) Brier score we report is $J^{-1} \sum_{j=1}^{J} B(q_{0.5}(\{y_{(j)i}\}_{i=1}^{n_{\text{test}}}) \mid \widehat{\boldsymbol{b}}_{(j)})$, where $q_{0.5}$ is the median of its set-valued argument. We also recorded Brier scores based on the 25th and 75th percentiles, but found their performances to be similar to the 50th.

# D Proof of Theorems 1 and 2

In this section, we prove the theoretical results in Section 4 of main manuscript. We will first list the assumptions under which we establish the asymptotic distribution of $\widehat{\boldsymbol{B}}_{r_*}$.

**(A1)** The data are independent and identically distributed from the Cox proportional hazards model described in Section 1 and $\text{rank}(\boldsymbol{B}_*) = r_*$.

**(A2)** For all $J$ populations, the data are collected on the finite time interval $[0, L]$ only.

**(A3)** For each $j \in [J]$ support of $\boldsymbol{x}_{(j)}$, $\mathcal{X}_{(j)}$, is a subset of $\mathbb{R}^p$ and $\max_{j\in[J]} \sup_{\mathbf{x}\in\mathcal{X}_{(j)}} \|\mathbf{x}\|_\infty \leq K$, for some finite $K \in \mathbb{R}$.

**(A4)** For each $j \in [J]$, we assume that there exists a finite constant $\tau_{(j)} \in (0, \infty)$ such that $\mathbb{P}(C_{(j)} \leq \tau_{(j)}) = 1$, $\mathbb{P}(C_{(j)} = \tau_{(j)}) > 0$ and $\mathbb{P}(T_{(j)} > \tau_{(j)}) > 0$.

**(A5)** For each $j \in [J]$, $n_{(j)}/n \to \kappa_{(j)} \geq \delta > 0$ for some $\delta > 0$ where $\sum_{j=1}^{J} \kappa_{(j)} = 1$.

**(A6)** The data from each of the $J$ populations are generated independently.

**(A7)** For all $j \in [J]$ and $s \in [0, L]$, assume that there exist $s \mapsto \boldsymbol{D}_{*(j)}(s)$ such that

(a) $n_{(j)}^{-1} \sum_{i\leq n_{(j)}} Y_{(j)i}(s) \exp(\boldsymbol{b}_{*(j)}^\top \boldsymbol{x}_{(j)i})(\boldsymbol{x}_{(j)i} - \bar{\boldsymbol{x}}_{(j)}(s))(\boldsymbol{x}_{(j)i} - \bar{\boldsymbol{x}}_{(j)}(s))^\top \xrightarrow{p} \boldsymbol{D}_{*(j)}(s)$ pointwise,

(b) $\int_0^L \boldsymbol{D}_{*(j)}(s) dH_{(j)}(s)$ is positive definite, where $H_{(j)}(s)$ is the cumulative baseline hazard for the $j$th population evaluated at $s$.

The above assumptions warrant some discussion. Assumption **(A1)** ensures that the model is well specified. Assumption **(A2)** can be easily be relaxed to allow for different finite intervals for each population. Assumptions **(A3)**, **(A4)**, and **(A7)** ensure that each of the $J$ populations have finite information. These assumptions are borrowed from Hjort and Pollard (2011) who studied the Cox model in a single population. They are weaker than those of Andersen and Gill (1982), e.g., Andersen and Gill (1982) require the convergence in assumption **(A7)** to be uniform in $s$ and for every $\boldsymbol{b}$ in the neighborhood of $\boldsymbol{b}_{*(j)}$, while **(A7)**(a) requires only pointwise convergence and only at the true regression coefficient. The bounded covariates assumption in **(A3)** can be relaxed with some careful calculations. Assumption **(A5)** ensures that each population is "equally" represented in the data. Assumption **(A6)** allows us to establish the joint distribution of unconstrained maximum likelihood estimators (MLEs) for each of the $J$ population.

With these assumptions in hand, we establish our first intermediate result on the consistency of $\widehat{\boldsymbol{B}}_{r_*}$.

**Lemma D.1** (Consistency)**.** *Under assumptions **(A1)** – **(A6)**, the constrained MLE is consistent, i.e.,* $\|\text{vec}(\widehat{\boldsymbol{B}}_{r_*} - \boldsymbol{B}_*)\|_2 = o_p(1)$.

*Proof.* Note that a Wald-type proof of consistency of the (constrained) maximum likelihood estimator is not possible because of the discontinuity of the profiled likelihoods. Instead, we will use techniques discussed in Section 5.3.2 of Van der Vaart (2002) to prove consistency of the MLE for a single population. Because the rank constraint requires that the MLE depend on data from all $J$ populations, the results from Van der Vaart (2002) cannot be applied to each of the populations separately.

Let $\widehat{H}_{(j)}$ be the cumulative baseline hazard function corresponding to the estimated baseline hazard $\widehat{h}_{(j)0}$.

For any bounded perturbation $\lambda_{(j)}$, let us define $d\widehat{H}_{(j)t_j} = (1 + t_j\lambda_{(j)})d\widehat{H}_{(j)}$. The joint likelihood at $(\widehat{\boldsymbol{B}}_{r_*}, \widehat{H}_{(1)t_1}, \ldots, \widehat{H}_{(J)t_J})$ viewed as a function of $t = (t_1, \ldots, t_J)$ must be maximized at $t = 0$ with $\widehat{\boldsymbol{B}}_{r_*}$ kept fixed. Thus we have $J$ stationary equations, for each $j \in [J]$,

$$\mathbb{P}_{n_{(j)}}\delta_{(j)}\lambda_{(j)}(y_{(j)}) = \int [\mathbb{P}_{n_{(j)}} \exp(\widehat{\boldsymbol{b}}_{(j)}^\top \boldsymbol{x}_{(j)})\mathbf{1}_{s\le y_{(j)}}]\lambda_{(j)}(s)d\widehat{H}_{(j)}(s),$$

where for every $j \in [J]$ and function $f : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$, we define

$$\mathbb{P}_{n_{(j)}} f := n_{(j)}^{-1} \sum_{i=1}^{n_{(j)}} f(y_{(j)i}, \boldsymbol{x}_{(j)i}).$$

By replacing $\lambda_{(j)}(s)$ by $\lambda_{(j)}(s)/[\mathbb{P}_{n_{(j)}} \exp(\widehat{\boldsymbol{b}}_{(j)}^\top \boldsymbol{x}_{(j)})\mathbf{1}_{s\le y_{(j)}}]$, we get for all $j \in [J]$,

$$\int \lambda_{(j)}(s)d\widehat{H}_{(j)}(s) = \mathbb{P}_{n_{(j)}} \frac{\delta_{(j)}\lambda_{(j)}(y_{(j)})}{\widehat{M}_{n(j)}(y_{(j)})}, \quad \text{where} \quad \widehat{M}_{n(j)}(s) := \mathbb{P}_{n_{(j)}} \exp(\widehat{\boldsymbol{b}}_{(j)}^\top \boldsymbol{x}_{(j)})\mathbf{1}_{s\le y_{(j)}}.$$

Note that the likelihood at $(\widehat{\boldsymbol{B}}_{r_*}, \widehat{H}_{(1)t_1}, \ldots, \widehat{H}_{(J)t_J})$ is not comparable to the likelihood at $(\boldsymbol{B}_*, H_{*(1)}, \ldots, H_{*(J)})$, where for $j \in [J]$, $H_{*(j)}$ is cumulative hazard function corresponding to $h_{*(j)0}$. This is because when maximizing the joint likelihood, the $\widehat{H}_{(j)t_j}$ are restricted to have discontinuities at the data points while $H_{*(j)}$ can be arbitrary. To overcome this, let us define an intermediate quantity $\widetilde{H}_{(j)}$ for each $j \in [J]$ that satisfies

$$\int \lambda_{(j)}(s)d\widetilde{H}_{(j)}(s) = \mathbb{P}_{n_{(j)}} \frac{\delta_{(j)}\lambda_{(j)}(y_{(j)})}{M_{(j)}(y_{(j)})}, \quad \text{where} \quad M_{(j)}(s) := P_{(j)} \exp(\boldsymbol{b}_{*(j)}^\top \boldsymbol{x}_{(j)})\mathbf{1}_{s\le y_{(j)}},$$

where for every $j \in [J]$ and function $f : \mathbb{R} \times \mathcal{X}_{(j)} \to \mathbb{R}$, we define

$$P_{(j)} f := \int f(y, \boldsymbol{x})dP_{(j)}(y, \boldsymbol{x}),$$

where $P_{(j)}$ denotes the joint distribution of $(y, \boldsymbol{x})$ for the $j$th population. By assumption **(A4)**, we have that

$$M_{(j)}(s) := P_{(j)} \exp(\boldsymbol{b}_{*(j)}^\top \boldsymbol{x}_{(j)})\mathbf{1}_{s\le y_{(j)}} \ge P_{(j)} \exp(\boldsymbol{b}_{*(j)}^\top \boldsymbol{x}_{(j)})\mathbf{1}_{\tau_{(j)}\le y_{(j)}}$$

is strictly bounded away from zero. Thus if $\lambda_{(j)}$ varies over a Glivenko-Cantelli class then $\int \lambda_{(j)}(s)d\widetilde{H}_{(j)} \to \int \lambda_{(j)}(s)dH_{*(j)}$ uniformly over $\lambda_{(j)}(s)$.

Now we will compare the likelihood at $(\widehat{\boldsymbol{B}}_{r_*}, \widehat{H}_{(1)}, \ldots, \widehat{H}_{(J)})$ to the likelihood at $(\boldsymbol{B}_*, \widetilde{H}_{(1)}, \ldots, \widetilde{H}_{(J)})$. Note that this is doable because both $\widehat{H}_{(j)}$ and $\widetilde{H}_{(j)}$ have point masses at the observed data and both $\widehat{\boldsymbol{B}}_{r_*}$ and $\boldsymbol{B}_*$ belong to the set of rank $r_*$ matrices. Comparing the likelihood, for each $j \in [J]$, we get

$$(\widehat{\boldsymbol{b}}_{(j)} - \boldsymbol{b}_{*(j)})^\top \mathbb{P}_{n_{(j)}} \boldsymbol{x}_{(j)}\delta_{(j)} - \mathbb{P}_{n_{(j)}} \left( e^{\widehat{\boldsymbol{b}}_{(j)}^\top \boldsymbol{x}_{(j)}} \widehat{H}_{(j)}(y_{(j)}) - e^{\boldsymbol{b}_{*(j)}^\top \boldsymbol{x}_{(j)}} \widetilde{H}_{(j)}(y_{(j)}) \right) + \mathbb{P}_{n_{(j)}} \delta_{(j)} \frac{M_{(j)}(y_{(j)})}{\widehat{M}_{n(j)}(y_{(j)})} \ge 0. \tag{14}$$

As $\|\widehat{\boldsymbol{b}}_{(j)}\|_2 \leq \|\widehat{\boldsymbol{B}}_{r_*}\|_F \leq M$ (see Section 4 of the main manuscript), by multiple applications of the Glivenko-Cantelli theorem, we will show that for all $j \in [J]$ and almost all $\omega$, there exists $\boldsymbol{b}_{\infty(j)}(\omega)$ and $H_{\infty(j)}(\omega)(\cdot)$ (a non-decreasing, cadlag function $H_{\infty(j)}(\omega) : [0,\infty) \to \mathbb{R}^+$ with $H_{\infty(j)}(\omega)(0) = 0$) such that along a subsequence $(\widehat{\boldsymbol{b}}_{(j)}(\omega), \widehat{H}_{(j)}(\omega)) \to (\boldsymbol{b}_{\infty(j)}(\omega), H_{\infty(j)}(\omega))$ and

$$(\boldsymbol{b}_{\infty(j)} - \boldsymbol{b}_{*(j)})^\top P_{(j)} \boldsymbol{x}_{(j)} \delta_{(j)} - P_{(j)} \left( e^{\boldsymbol{b}_{\infty(j)}^\top \boldsymbol{x}_{(j)}} H_{\infty(j)}(y_{(j)}) - e^{\boldsymbol{b}_{*(j)}^\top \boldsymbol{x}_{(j)}} H_{*(j)}(y_{(j)}) \right) + P_{(j)} \delta_{(j)} \frac{M_{(j)}(y_{(j)})}{M_{\infty(j)}(y_{(j)})} \geq 0, \tag{15}$$

where

$$M_{\infty(j)}(s) := P_{(j)} e^{\boldsymbol{b}_{\infty(j)}^\top \boldsymbol{x}_{(j)}} \mathbf{1}_{s \leq y_{(j)}} \quad \text{and} \quad \int \lambda_{(j)} dH_{\infty(j)} = P_{(j)} \frac{\delta_{(j)} \lambda_{(j)}(y_{(j)})}{M_{\infty(j)}(y_{(j)})}.$$

Now note that by a perturbation similar to the one in the beginning of the proof, we have that

$$\int \lambda_{(j)}(s) dH_{*(j)}(s) = P_{(j)} \frac{\delta_{(j)} \lambda_{(j)}(y_{(j)})}{M_{(j)}(y_{(j)})}.$$

Let $l(\boldsymbol{b}_{n(j)}, H_{n(j)})$ be the likelihood evaluated at $(\boldsymbol{b}_{n(j)}, H_{n(j)})$. By observing that $M_{(j)}/M_{\infty(j)}(y_{(j)}) = dH_{\infty(j)}/dH_{*(j)}$, we see that (15) implies that

$$P_{(j)} \log\{ l(\boldsymbol{b}_{\infty(j)}, H_{\infty(j)}) / l(\boldsymbol{b}_{*(j)}, H_{*(j)}) \} \leq 0,$$

which by uniqueness of $(\boldsymbol{b}_{*(j)}, H_{*(j)})$ implies that $(\boldsymbol{b}_{\infty(j)}, H_{\infty(j)}) = (\boldsymbol{b}_{*(j)}, H_{*(j)})$ almost surely.

The proof will be complete if we can show that (14) implies (15) for each of the populations separately. This is done on page 392 of Section 5.3.2 of Van der Vaart (2002). This last step uses the fact that $\chi_{(j)}$ is a bounded set (assumption **(A3)**) and the fact that uniformly bounded cadlag functions on bounded support are Glivenko-Cantelli. □

## D.1 Proof of Theorem 1

Next, we prove the asymptotic normality of $\mathrm{vec}(\widehat{\boldsymbol{B}}_{r_*} - \boldsymbol{B}_*)$. For the remainder of this section, we take $r = r_*$. For every $j \in [J]$, define

$$R_{(j)}(s, \boldsymbol{b}) := \sum_{i=1}^{n_{(j)}} \mathbf{1}(y_{(j)i} \geq s) \exp\big(\boldsymbol{x}_{(j)i}^\top \boldsymbol{b}\big) \quad \text{and} \quad dN_{(j)i}(s) = \mathbf{1}\big(y_{(j)i} \in [s, s+ds], \delta_{(j)i} = 1\big). \tag{16}$$

Recall that the partial log likelihood for the data is

$$\begin{aligned} \boldsymbol{\mathcal{L}}(\boldsymbol{B}) &:= \sum_{j=1}^{J} \sum_{i=1}^{n_{(j)}} \delta_{(j)i} \left[ \boldsymbol{x}_{(j)i}^\top \boldsymbol{b}_{(j)} - \log \Big\{ \sum_{k \in \mathcal{R}_{(j)i}} \exp\big(\boldsymbol{x}_{(j)k}^\top \boldsymbol{b}_{(j)}\big) \Big\} \right] \\ &= \sum_{j=1}^{J} \sum_{i=1}^{n_{(j)}} \int_0^L \left[ \boldsymbol{x}_{(j)i}^\top \boldsymbol{b}_{(j)} - \log R_j(s, \boldsymbol{b}_{(j)}) \right] dN_{(j)i}(s), \end{aligned} \tag{17}$$

where the risk set $\mathcal{R}_{(j)i}$ is defined Section 1 of the manuscript. Let us now define a number of important quantities. For two (compatible) matrices $\boldsymbol{J}$ and $\boldsymbol{K}$, let the notation $\boldsymbol{J}_{[,1:r]}$ denote the first $r$ columns of $\boldsymbol{J}$, and let $[\boldsymbol{J}, \boldsymbol{K}]$ denote the matrix which concatenates $\boldsymbol{J}$ and $\boldsymbol{K}$ by columns. Let $e_j \in \mathbb{R}^r$ denote the $j$th basis vector for $j \in [r]$. Let $\widetilde{\boldsymbol{U}} \in \mathbb{R}^{p\times r}$ and $\widetilde{\boldsymbol{V}} \in \mathbb{R}^{(J-r)\times r}$ be two unique matrices such that

$$\boldsymbol{B}_* = [\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{V}}^\top, \widetilde{\boldsymbol{U}}].$$

For any vector $\boldsymbol{\beta} \in \mathbb{R}^{pr+Jr}$, we use the shorthand

$$\boldsymbol{\beta}_{\widetilde{\boldsymbol{U}}} := \boldsymbol{\beta}_{[1:pr]}, \quad \text{and for } j \in [J], \quad \boldsymbol{\beta}_{\widetilde{\boldsymbol{V}}_{(j)}} := \boldsymbol{\beta}_{[(pr+(j-1)r+1):(pr+jr)]}. \tag{18}$$

Further, let $\boldsymbol{\alpha} := (\mathrm{vec}(\widetilde{\boldsymbol{U}})^\top, \mathrm{vec}([\widetilde{\boldsymbol{V}}^\top, I_r])^\top)^\top \in \mathbb{R}^{pr+Jr}$ so that we may write, for example,

$$\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(j)}} = \begin{cases} \widetilde{\boldsymbol{V}}_{[j,]} & \text{for } j \le J-r, \\ e_{j-(J-r)} & \text{for } j > J-r. \end{cases}$$

Finally, let us define the matrices $\boldsymbol{T}$ and $\boldsymbol{T}_1$ as

$$\begin{aligned} \boldsymbol{T} &:= \left[\left(\begin{bmatrix} \widetilde{\boldsymbol{V}} \\ I_r \end{bmatrix} \otimes I_p\right), I_J \otimes \widetilde{\boldsymbol{U}}\right] \in \mathbb{R}^{pJ\times(pr+Jr)}, \\ \boldsymbol{T}_1 &:= \boldsymbol{T}_{[,1:(pr+r(J-r))]} = \left[\left(\begin{bmatrix} \widetilde{\boldsymbol{V}} \\ I_r \end{bmatrix} \otimes I_p\right), (I_J)_{[,1:(J-r)]} \otimes \widetilde{\boldsymbol{U}}\right] \in \mathbb{R}^{pJ\times(pr+r(J-r))}. \end{aligned} \tag{19}$$

To simplify our notation, we re-express the partial log-likelihood in (17) as $\mathcal{L} : \mathbb{R}^{pr+Jr} \to \mathbb{R}$ with

$$\mathcal{L}(\boldsymbol{\beta}) := \sum_{j=1}^{J}\sum_{i=1}^{n_{(j)}} \int_0^L \Big[\boldsymbol{x}_{(j)i}^\top \boldsymbol{b}_{(j)}(\boldsymbol{\beta}) - \log R_j\{s, \boldsymbol{b}_{(j)}(\boldsymbol{\beta})\}\Big] dN_{(j)i}(s),$$

where for any $\boldsymbol{\beta} \in \mathbb{R}^{pr+rJ}$, with a slight abuse of notation, we redefine $\boldsymbol{b}_{(j)} : \mathbb{R}^{pr+rJ} \to \mathbb{R}^p$,

$$\boldsymbol{b}_{(j)}(\boldsymbol{\beta}) := (\boldsymbol{\beta}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\beta}_{\widetilde{\boldsymbol{U}}} \in \mathbb{R}^p, \tag{20}$$

with $\boldsymbol{\beta}_{\widetilde{\boldsymbol{V}}_{(j)}}$ and $\boldsymbol{\beta}_{\boldsymbol{U}}$ as defined in (18). In the above notation $\boldsymbol{b}_{(j)}(\boldsymbol{\alpha}) = \boldsymbol{b}_{*(j)}$ for all $j \in [J]$. Next, define

$$\bar{\boldsymbol{x}}_{(j)}(s) := \sum_{i=1}^{n_{(j)}} \boldsymbol{x}_{(j)i}\pi_{(j)i}(s), \qquad \mathbf{F}_{(j)}(s) := \sum_{i=1}^{n_{(j)}} \pi_{(j)i}(s)\big(\boldsymbol{x}_{(j)i} - \bar{\boldsymbol{x}}_{(j)}(s)\big)\big(\boldsymbol{x}_{(j)i} - \bar{\boldsymbol{x}}_{(j)}(s)\big)^\top,$$

$$\text{and} \qquad \pi_{(j)i}(s) := \frac{\mathbf{1}(y_{(j)i} \ge s)\exp(\boldsymbol{x}_{(j)i}^\top \boldsymbol{b}_{*(j)})}{R_j(s, \boldsymbol{b}_{*(j)})}.$$

Letting

$$\mathcal{G} := \mathbb{R}^{pr+r(J-r)} \times \underbrace{\{0\} \times \ldots \times \{0\}}_{r^2}, \tag{21}$$

in Lemma E.1, we show that for any $\boldsymbol{\nu} \in \mathcal{G}$, we can write

$$\log R_{(j)}\{s, \boldsymbol{b}_{(j)}(\boldsymbol{\alpha}+\boldsymbol{\nu})\} - \log R_{(j)}\{s, \boldsymbol{b}_{(j)}(\boldsymbol{\alpha})\} = \bar{\boldsymbol{x}}_{(j)}(s)^\top \boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}) + \frac{1}{2}\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu})^\top \mathbf{F}_{(j)}(s)\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}) + r_{(j)}(\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}), s), \tag{22}$$

where $\boldsymbol{b}_{(j)}(\cdot)$ is as defined in (20) and

$$\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}) := \boldsymbol{b}_{(j)}(\boldsymbol{\alpha}+\boldsymbol{\nu}) - \boldsymbol{b}_{(j)}(\boldsymbol{\alpha}) = (\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\alpha}_{\widetilde{\boldsymbol{U}}} + (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}} + (\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}. \tag{23}$$

Note that for $j > J - r$, we have $\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}) = (\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}})_{[(jp-p+1):jp]}$ since in this case, $\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}} = \mathbf{0}_r$ (by (21)) and $\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(j)}} = e_{j-(J-r)}$. Thus we have that

$$\begin{aligned}
\mathcal{L}^*(\boldsymbol{\nu}) &:= \mathcal{L}\left(\boldsymbol{\alpha} + \boldsymbol{\nu}/\sqrt{n}\right) - \mathcal{L}(\boldsymbol{\alpha}) \\
&= \sum_{j=1}^{J}\sum_{i=1}^{n_{(j)}} \int_0^L (\boldsymbol{x}_{(j)i} - \bar{\boldsymbol{x}}_{(j)}(s))^\top \boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}/\sqrt{n}) dN_{(j)i}(s) \\
&\quad - \sum_{j=1}^{J}\sum_{i=1}^{n_{(j)}} \int_0^L \Big[\frac{1}{2}\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}/\sqrt{n})^\top \mathbf{F}_{(j)}(s)\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}/\sqrt{n}) - r_{(j)}\big(\boldsymbol{\gamma}_j(\boldsymbol{\alpha}, \boldsymbol{\nu}/\sqrt{n}), s\big)\Big] dN_{(j)i}(s) \\
&= \sum_{j=1}^{J} \sqrt{n}\boldsymbol{a}_{(j)}^\top \boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}/\sqrt{n}) - \sum_{j=1}^{J} \frac{n}{2}\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}/\sqrt{n})^\top \boldsymbol{D}_{(j)}\boldsymbol{\gamma}_j(\boldsymbol{\alpha}, \boldsymbol{\nu}/\sqrt{n}) \\
&\quad - \sum_{j=1}^{J}\sum_{i=1}^{n_{(j)}} \int_0^L r_{(j)}\big(\boldsymbol{\gamma}_j(\boldsymbol{\alpha}, \boldsymbol{\nu}/\sqrt{n}), s\big) dN_{(j)i}(s) \\
&= \sum_{j=1}^{J} \boldsymbol{a}_{(j)}^\top (\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\alpha}_{\widetilde{\boldsymbol{U}}} + \sum_{j=1}^{J} \boldsymbol{a}_{(j)}^\top (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}} + n^{-1/2}\sum_{j=1}^{J} \boldsymbol{a}_{(j)}^\top (\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}} \\
&\quad - \frac{1}{2}\sum_{j=1}^{J} \big((\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\alpha}_{\widetilde{\boldsymbol{U}}} + (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}\big)^\top \boldsymbol{D}_{(j)}\big((\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\alpha}_{\widetilde{\boldsymbol{U}}} + (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}\big) \\
&\quad - \frac{1}{\sqrt{n}}\sum_{j=1}^{J} \big((\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}\big)^\top \boldsymbol{D}_{(j)}\big((\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\alpha}_{\widetilde{\boldsymbol{U}}} + (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}\big) \\
&\quad - \frac{1}{2n}\sum_{j=1}^{J} \big((\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}\big)^\top \boldsymbol{D}_{(j)}\big((\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}\big) \\
&\quad - \sum_{j=1}^{J}\sum_{i=1}^{n_{(j)}} \int_0^L r_{(j)}\big(\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}/\sqrt{n}), s\big) dN_{(j)i}(s),
\end{aligned}$$

where

$$\boldsymbol{a}_{(j)} := n^{-1/2}\sum_{i=1}^{n_{(j)}} \int_0^L (\boldsymbol{x}_{(j)i} - \bar{\boldsymbol{x}}_{(j)}(s)) dN_{(j)i}(s) \in \mathbb{R}^p, \text{ and } \boldsymbol{D}_{(j)} := n^{-1}\sum_{i=1}^{n_{(j)}} \int_0^L \mathbf{F}_{(j)}(s) dN_{(j)i}(s) \in \mathbb{R}^{p\times p}. \tag{24}$$

Note that the scaling above is with respect to the $n = \sum_{j=1}^{J} n_{(j)}$. Defining

$$
\begin{aligned}
r_n(\boldsymbol{\nu}) := n^{-1/2} \sum_{j=1}^{J} \boldsymbol{a}_{(j)}^\top (\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p) \boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}} - \frac{1}{2n} \sum_{j=1}^{J} \big((\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p) \boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}\big)^\top \boldsymbol{D}_{(j)} \big((\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_j}^\top \otimes I_p) \boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}\big) \\
- \sum_{j=1}^{J} \sum_{i=1}^{n_{(j)}} \int_0^L r_{(j)}\big(\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}/\sqrt{n}), s\big) dN_{(j)i}(s) \qquad (25) \\
- \frac{1}{\sqrt{n}} \sum_{j=1}^{J} \big((\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p) \boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}\big)^\top \boldsymbol{D}_{(j)} \big((\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_j}^\top \otimes I_p) \boldsymbol{\alpha}_{\widetilde{\boldsymbol{U}}} + (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p) \boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}\big),
\end{aligned}
$$

observe that $\mathcal{L}^*(\boldsymbol{\nu})$ is a quadratic in $\boldsymbol{\nu}$ with the following expression:

$$
\mathcal{L}^*(\boldsymbol{\nu}) = \widetilde{\boldsymbol{A}} \boldsymbol{\nu} - \frac{1}{2} \boldsymbol{\nu}^\top \boldsymbol{Q} \boldsymbol{\nu} + r_n(\boldsymbol{\nu}), \qquad (26)
$$

where

$$
\widetilde{\boldsymbol{A}} := \begin{pmatrix} \sum_{j=1}^{J} \boldsymbol{a}_{(j)}^\top (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p) \\ \boldsymbol{a}_{(1)}^\top \mathrm{vec}^{-1}(\boldsymbol{\alpha}_{\widetilde{\boldsymbol{U}}}) \\ \boldsymbol{a}_{(2)}^\top \mathrm{vec}^{-1}(\boldsymbol{\alpha}_{\widetilde{\boldsymbol{U}}}) \\ \vdots \\ \boldsymbol{a}_{(J)}^\top \mathrm{vec}^{-1}(\boldsymbol{\alpha}_{\widetilde{\boldsymbol{U}}}) \end{pmatrix}^\top = (\boldsymbol{a}_{(1)}^\top, \boldsymbol{a}_{(2)}^\top, \cdots, \boldsymbol{a}_{(J)}^\top) \underbrace{\left[ \left( \begin{bmatrix} \widetilde{\boldsymbol{V}} \\ I_r \end{bmatrix} \otimes I_p \right), I_J \otimes \widetilde{\boldsymbol{U}} \right]}_{\boldsymbol{T}} = \boldsymbol{A}^\top \boldsymbol{T},
$$

with

$$
\boldsymbol{A}^\top := (\boldsymbol{a}_{(1)}^\top, \boldsymbol{a}_{(2)}^\top, \cdots, \boldsymbol{a}_{(J)}^\top), \qquad (27)
$$

and $\boldsymbol{T}$ as defined in (19). Moreover, $\boldsymbol{Q}$ can be expressed

$$
\boldsymbol{Q} := \begin{pmatrix}
\sum_{j=1}^{J} (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)^\top \boldsymbol{D}_{(j)} (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p) & (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(1)}}^\top \otimes I_p)^\top \boldsymbol{D}_{(1)} \widetilde{\boldsymbol{U}} & \cdots & \cdots & (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(J)}}^\top \otimes I_p)^\top \boldsymbol{D}_{(J)} \widetilde{\boldsymbol{U}} \\
\widetilde{\boldsymbol{U}}^\top \boldsymbol{D}_{(1)} (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(1)}}^\top \otimes I_p) & \widetilde{\boldsymbol{U}}^\top \boldsymbol{D}_{(1)} \widetilde{\boldsymbol{U}} & 0 & \cdots & 0 \\
\widetilde{\boldsymbol{U}}^\top \boldsymbol{D}_{(2)} (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(2)}}^\top \otimes I_p) & 0 & \widetilde{\boldsymbol{U}}^\top \boldsymbol{D}_{(2)} \widetilde{\boldsymbol{U}} & \cdots & \vdots \\
\vdots & \vdots & 0 & \ddots & 0 \\
\widetilde{\boldsymbol{U}}^\top \boldsymbol{D}_{(J)} (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(J)}}^\top \otimes I_p) & 0 & \cdots & 0 & \widetilde{\boldsymbol{U}}^\top \boldsymbol{D}_{(J)} \widetilde{\boldsymbol{U}}
\end{pmatrix}.
$$

and

$$
\boldsymbol{D} := \mathrm{BlockDiag}\{\boldsymbol{D}_{(j)}\}_{j=1}^{J} \in \mathbb{R}^{pJ \times pJ}, \qquad (28)
$$

so that we may alternatively write $\boldsymbol{Q}$ more compactly as

$$
\boldsymbol{Q} = \boldsymbol{T}^\top \boldsymbol{D} \boldsymbol{T}.
$$

Since $\boldsymbol{\nu} \in \mathcal{G}$ has $r^2$ constrained coordinates, we will now rewrite the (centered) partial likelihood in terms of the *free* coordinates, which we denote by $\boldsymbol{\eta} \in \mathbb{R}^{pr + r(J-r)}$, i.e., $\boldsymbol{\nu}^\top =$

$(\boldsymbol{\eta}^\top, \mathbf{0}_{r^2}^\top)$. Thus writing (26) as function of $\boldsymbol{\eta}$, we get

$$\begin{aligned}\widetilde{\mathcal{L}}(\boldsymbol{\eta}) := \mathcal{L}^*(\boldsymbol{\nu}) &= \boldsymbol{A}^\top \boldsymbol{T}\boldsymbol{\nu} - \frac{1}{2}\boldsymbol{\nu}^\top(\boldsymbol{T}^\top \boldsymbol{D}\boldsymbol{T})\boldsymbol{\nu} + r_n(\boldsymbol{\nu}) \\ &= \boldsymbol{A}^\top \boldsymbol{T}_1\boldsymbol{\eta} - \frac{1}{2}\boldsymbol{\eta}^\top(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)\boldsymbol{\eta} + r_n\{(\boldsymbol{\eta}^\top, \mathbf{0}_{r^2}^\top)^\top\},\end{aligned} \tag{29}$$

where $\boldsymbol{A}, \boldsymbol{D}$, and $\boldsymbol{T}_1$ are as defined in (27), (28), and (19), respectively. Letting $\widehat{\boldsymbol{\eta}}$ be the maximizer of $\widetilde{\mathcal{L}}(\cdot)$ (note that this is unique), (26) and Lemma E.1 imply

$$\widetilde{\mathcal{L}}\left(\mathbf{0}_{pr+r(J-r)}\right) = 0 \leq \boldsymbol{A}^\top \boldsymbol{T}_1\widehat{\boldsymbol{\eta}} - \frac{1}{2}\widehat{\boldsymbol{\eta}}^\top(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)\widehat{\boldsymbol{\eta}} + r_n\{(\widehat{\boldsymbol{\eta}}^\top, \mathbf{0}_{r^2}^\top)^\top\} = \widetilde{\mathcal{L}}(\widehat{\boldsymbol{\eta}}). \tag{30}$$

We will now show that $\|\widehat{\boldsymbol{\eta}}\|_2 = O_p(1)$. Recall that by consistency of the rank constrained MLE (Lemma D.1) we have that

$$\frac{\|\widehat{\boldsymbol{\eta}}\|_2}{\sqrt{n}} = o_p(1).$$

Thus by (37) (Lemma E.1), we have that

$$\begin{aligned} r_n\{(\widehat{\boldsymbol{\eta}}^\top, \mathbf{0}_{r^2}^\top)^\top\} &= O_p\left(\frac{\|\widehat{\boldsymbol{\eta}}\|_2^2}{\sqrt{n}} + \frac{\|\widehat{\boldsymbol{\eta}}\|_2^4}{n} + \frac{\|\widehat{\boldsymbol{\eta}}\|_2^3}{\sqrt{n}}\left[1 + \left(\frac{\|\widehat{\boldsymbol{\eta}}\|_2}{\sqrt{n}}\right)^3\right]\right) \\ &= O_p\left(\|\widehat{\boldsymbol{\eta}}\|_2^2\left[n^{-1/2} + \frac{\|\widehat{\boldsymbol{\eta}}\|_2^2}{n} + \frac{\|\widehat{\boldsymbol{\eta}}\|_2}{\sqrt{n}}\left[1 + \left(\frac{\|\widehat{\boldsymbol{\eta}}\|_2}{\sqrt{n}}\right)^3\right]\right]\right) = o_p(\|\widehat{\boldsymbol{\eta}}\|_2^2).\end{aligned} \tag{31}$$

Suppose $\|\widehat{\boldsymbol{\eta}}\|_2 \to \infty$. Dividing (30) by $(1 + \|\widehat{\boldsymbol{\eta}}\|_2)^2$, by the above display we get

$$0 \leq \boldsymbol{A}^\top \boldsymbol{T}_1 \frac{\widehat{\boldsymbol{\eta}}}{(1 + \|\widehat{\boldsymbol{\eta}}\|_2)^2} - \frac{1}{2}\frac{(\boldsymbol{T}_1\widehat{\boldsymbol{\eta}})^\top \boldsymbol{D}(\boldsymbol{T}_1\widehat{\boldsymbol{\eta}})}{(1 + \|\widehat{\boldsymbol{\eta}}\|_2)^2} + o_p\left(\frac{\|\widehat{\boldsymbol{\eta}}\|_2^2}{(1 + \|\widehat{\boldsymbol{\eta}}\|_2)^2}\right).$$

As $\|\widehat{\boldsymbol{\eta}}\|_2 \to \infty$, we have $\|\widehat{\boldsymbol{\eta}}\|_2/(1 + \|\widehat{\boldsymbol{\eta}}\|_2)^2 = o_p(1)$ and thus $0 \leq \widehat{\boldsymbol{\eta}}^\top(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)\widehat{\boldsymbol{\eta}}/(1 + \|\widehat{\boldsymbol{\eta}}\|_2)^2 \leq o_p(1)$, but this is a contradiction as the smallest eigenvalue of $\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1$ is bounded away from 0. Thus we conclude that $\|\widehat{\boldsymbol{\eta}}\|_2 = O_p(1)$, which proves $\sqrt{n}$-consistency of $\widehat{\boldsymbol{\eta}}$. Hence $r_n\{(\widehat{\boldsymbol{\eta}}^\top, \mathbf{0}_{r^2}^\top)^\top\} = o_p(1)$, thus we have

$$\widetilde{\mathcal{L}}(\widehat{\boldsymbol{\eta}}) = \boldsymbol{A}^\top \boldsymbol{T}_1\widehat{\boldsymbol{\eta}} - \frac{1}{2}\widehat{\boldsymbol{\eta}}^\top \boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1\widehat{\boldsymbol{\eta}} + o_p(1).$$

Below, we will establish the following the three results:

**(R1)** $\boldsymbol{T}_1\widehat{\boldsymbol{\eta}} = \boldsymbol{T}_1(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1^\top \boldsymbol{A} + o_p(1)$.

**(R2)** $\sqrt{n}\{\mathrm{vec}(\boldsymbol{B}_{\boldsymbol{\alpha}+\widehat{\boldsymbol{\nu}}/\sqrt{n}} - \boldsymbol{B}_*)\} = \boldsymbol{T}_1\widehat{\boldsymbol{\eta}} + O_p(n^{-1/2}\|\widehat{\boldsymbol{\eta}}\|_2^2) = \boldsymbol{T}_1\widehat{\boldsymbol{\eta}} + O_p(n^{-1/2})$.

**(R3)** For every pair $(\boldsymbol{U}, \boldsymbol{V})$ such that $\boldsymbol{B}_* = \boldsymbol{U}\boldsymbol{V}^\top$, we have

$$\boldsymbol{T}_1(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1 = \boldsymbol{T}_{(\boldsymbol{U},\boldsymbol{V})}(\boldsymbol{T}_{(\boldsymbol{U},\boldsymbol{V})}^\top \boldsymbol{D}\boldsymbol{T}_{(\boldsymbol{U},\boldsymbol{V})})^+\boldsymbol{T}_{(\boldsymbol{U},\boldsymbol{V})},$$

where

$$\boldsymbol{T}_{(\boldsymbol{U},\boldsymbol{V})} := [\boldsymbol{V} \otimes I_p, I_J \otimes \boldsymbol{U}]. \tag{32}$$

The main result then follows from an application of Slutsky's theorem in conjunction with **(R1)**, **(R2)**, **(R3)**, and the asymptotic normality of $\boldsymbol{A}$ in Lemma E.2.

**Proof of (R1).**

Defining $\boldsymbol{H} := \boldsymbol{T}_1(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1^\top \boldsymbol{A}$, we get $\boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{T}_1 = \boldsymbol{A}^\top \boldsymbol{T}_1(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1\boldsymbol{D}\boldsymbol{T}_1 = \boldsymbol{A}^\top \boldsymbol{T}_1$. By completing the square, it follows that

$$\begin{aligned}
\widetilde{\mathcal{L}}(\widehat{\boldsymbol{\eta}}) &= \boldsymbol{A}^\top \boldsymbol{T}_1\widehat{\boldsymbol{\eta}} - \frac{1}{2}\widehat{\boldsymbol{\eta}}^\top \boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1\widehat{\boldsymbol{\eta}} + o_p(1) \\
&= \boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{T}_1\widehat{\boldsymbol{\eta}} - \frac{1}{2}\widehat{\boldsymbol{\eta}}^\top \boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1\widehat{\boldsymbol{\eta}} + o_p(1) \\
&= \boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{T}_1\widehat{\boldsymbol{\eta}} - \frac{1}{2}\widehat{\boldsymbol{\eta}}^\top \boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1\widehat{\boldsymbol{\eta}} - \frac{1}{2}\boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{H} + \frac{1}{2}\boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{H} + o_p(1) \\
&= -\frac{1}{2}(\boldsymbol{T}_1\widehat{\boldsymbol{\eta}} - \boldsymbol{H})^\top \boldsymbol{D}(\boldsymbol{T}_1\widehat{\boldsymbol{\eta}} - \boldsymbol{H}) + \frac{1}{2}\boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{H} + o_p(1).
\end{aligned} \tag{33}$$

Next, let $\widetilde{\boldsymbol{\eta}} = (\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1^\top \boldsymbol{A}$, by (29), we have that

$$\widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\eta}}) = \boldsymbol{A}^\top \boldsymbol{T}\widetilde{\boldsymbol{\eta}} - \frac{1}{2}\widetilde{\boldsymbol{\eta}}^\top (\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)\widetilde{\boldsymbol{\eta}} + r_n\{(\widetilde{\boldsymbol{\eta}}^\top, \boldsymbol{0}_{r^2}^\top)^\top\}. \tag{34}$$

By Lemma E.2 we have that $\|\widetilde{\boldsymbol{\eta}}\|_2 = O_p(1)$. Thus (34) and (31) implies

$$\begin{aligned}
\widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\eta}}) &= \boldsymbol{A}^\top \boldsymbol{T}_1\widetilde{\boldsymbol{\eta}} - \frac{1}{2}\widetilde{\boldsymbol{\eta}}^\top \boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1\widetilde{\boldsymbol{\eta}} + o_p(1) \\
&= \boldsymbol{A}^\top \boldsymbol{T}_1(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1^\top \boldsymbol{A} - \frac{1}{2}\boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{H} + o_p(1) \\
&= \boldsymbol{A}^\top \boldsymbol{T}_1(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)^{-1}(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1^\top \boldsymbol{A} - \frac{1}{2}\boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{H} + o_p(1) \\
&= \boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{T}_1(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1^\top \boldsymbol{A} - \frac{1}{2}\boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{H} + o_p(1) \\
&= \boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{H} - \frac{1}{2}\boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{H} + o_p(1) = \frac{1}{2}\boldsymbol{H}^\top \boldsymbol{D}\boldsymbol{H} + o_p(1).
\end{aligned} \tag{35}$$

As $\boldsymbol{\alpha} + (\widetilde{\boldsymbol{\eta}}^\top/\sqrt{n}, \boldsymbol{0}_{r^2}^\top)^\top$ lives in the constrained parameter space and $\widehat{\boldsymbol{\eta}}$ is the rank constrained MLE, we have that $\widetilde{\mathcal{L}}(\widehat{\boldsymbol{\eta}}) \geq \widetilde{\mathcal{L}}(\widetilde{\boldsymbol{\eta}})$. Combining this with (33) and (35) we get

$$-\frac{1}{2}(\boldsymbol{T}_1\widehat{\boldsymbol{\eta}} - \boldsymbol{T}_1(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1^\top \boldsymbol{A})^\top \boldsymbol{D}(\boldsymbol{T}_1\widehat{\boldsymbol{\eta}} - \boldsymbol{T}_1(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1\boldsymbol{A}) + o_p(1) \geq 0.$$

**Proof of (R2).**

Defining $\widehat{\boldsymbol{\nu}} := (\widehat{\boldsymbol{\eta}}^\top, \boldsymbol{0}_{r^2}^\top)^\top$, We will now relate $\widehat{\boldsymbol{\nu}}$ to $\sqrt{n}\{\mathrm{vec}(\boldsymbol{B}_{\boldsymbol{\alpha}+\widehat{\boldsymbol{\nu}}/\sqrt{n}} - \boldsymbol{B}^*)\}$. Recall that

$$\boldsymbol{\alpha} := (\mathrm{vec}(\widetilde{\boldsymbol{U}})^\top, \mathrm{vec}([\widetilde{\boldsymbol{V}}^\top, I_r])^\top)^\top \text{ and } \boldsymbol{\alpha}+\boldsymbol{\nu}/\sqrt{n} = \{(\mathrm{vec}(\widetilde{\boldsymbol{U}})^\top, \mathrm{vec}([\widetilde{\boldsymbol{V}}^\top, I_r])^\top) + (\boldsymbol{\eta}^\top, \boldsymbol{0}_{r^2}^\top)/\sqrt{n}\}^\top.$$

Observe that

$$\begin{aligned}
\sqrt{n}(\boldsymbol{B}_{\boldsymbol{\alpha}+\widehat{\boldsymbol{\nu}}/\sqrt{n}} - \boldsymbol{B}^*) &= \sqrt{n}\Big\{(\widetilde{\boldsymbol{U}} + \mathrm{vec}^{-1}(\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{U}}})/\sqrt{n})[\widetilde{\boldsymbol{V}}^\top + \mathrm{vec}^{-1}(\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{V}}})/\sqrt{n}, I_r] - \widetilde{\boldsymbol{U}}[\widetilde{\boldsymbol{V}}^\top, I_r]\Big\} \\
&= \mathrm{vec}^{-1}(\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{U}}})[\widetilde{\boldsymbol{V}}^\top, I_r] + \widetilde{\boldsymbol{U}}[\mathrm{vec}^{-1}(\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{V}}}), \boldsymbol{0}_{r\times r}] + \mathrm{vec}^{-1}(\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{U}}})[\mathrm{vec}^{-1}(\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{V}}}), \boldsymbol{0}_{r\times r}]/\sqrt{n},
\end{aligned}$$

where, for example, $\mathrm{vec}^{-1}(\boldsymbol{\alpha}_{\widetilde{\boldsymbol{U}}}) = \widetilde{\boldsymbol{U}}$ and $\mathrm{vec}^{-1}(\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}}) = \widetilde{\boldsymbol{V}}^\top$. Notice

$$\mathrm{vec}\left\{\mathrm{vec}^{-1}(\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{U}}})[\widetilde{\boldsymbol{V}}^\top, I_r] + \widetilde{\boldsymbol{U}}[\mathrm{vec}^{-1}(\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{V}}}), \boldsymbol{0}_{r\times r}]\right\} = ([\widetilde{\boldsymbol{V}}^\top, I_r]^\top \otimes I_p)\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{U}}} + (I \otimes \widetilde{\boldsymbol{U}})(\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{V}}}^\top, \boldsymbol{0}_{r^2}^\top)^\top = \boldsymbol{T}_1\widehat{\boldsymbol{\eta}},$$

and thus, with $\boldsymbol{T}_1$ defined in (19) we have **(R2)**.

**Proof of (R3).**

Let us consider any pair $(\boldsymbol{U}, \boldsymbol{V})$ where $\boldsymbol{U} \in \mathbb{R}^{p\times r}$, $\boldsymbol{V} \in \mathbb{R}^{J\times r}$ such that $\boldsymbol{B}_* = \boldsymbol{U}\boldsymbol{V}^\top$. Note that $(\widetilde{\boldsymbol{U}}, [\widetilde{\boldsymbol{V}}^\top, I_r]^\top)$ is one such possible pair of $(\boldsymbol{U}, \boldsymbol{V})$. Recall $\boldsymbol{T}_{(\boldsymbol{U},\boldsymbol{V})} = [\boldsymbol{V} \otimes I_p, I_J \otimes \boldsymbol{U}]$ as defined in (32). It is easy to see that $\mathrm{col}(\boldsymbol{T}_{(\boldsymbol{U},\boldsymbol{V})}) = \mathrm{col}(\boldsymbol{T})$, i.e., the column space of $\boldsymbol{T}_{(\boldsymbol{U},\boldsymbol{V})}$ and $\boldsymbol{T}$ are the same for all $(\boldsymbol{U}, \boldsymbol{V})$ such that $\boldsymbol{B}_* = \boldsymbol{U}\boldsymbol{V}^\top$. Hence, it follows immediately that

$$\boldsymbol{T}_1(\boldsymbol{T}_1^\top \boldsymbol{D}\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1 = \boldsymbol{T}(\boldsymbol{T}^\top\boldsymbol{D}\boldsymbol{T})^{+}\boldsymbol{T} = \boldsymbol{T}_{(\boldsymbol{U},\boldsymbol{V})}(\boldsymbol{T}_{(\boldsymbol{U},\boldsymbol{V})}^\top \boldsymbol{D}\boldsymbol{T}_{(\boldsymbol{U},\boldsymbol{V})})^{+}\boldsymbol{T}_{(\boldsymbol{U},\bar{\boldsymbol{V}})}.$$

## D.2 Proof of Theorem 2

To prove that the asymptotic variance of $\widehat{\boldsymbol{B}}_{r_*}$ is less than that of the unconstrained MLE, notice that

$$\begin{aligned}
&\mathrm{avar}[\sqrt{n}\{\mathrm{vec}(\bar{\boldsymbol{B}} - \boldsymbol{B}_*)\}] - \mathrm{avar}[\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{B}}_{r_*} - \boldsymbol{B}_*)\}] \\
&\qquad = \boldsymbol{D}_*^{-1} - \boldsymbol{T}_1(\boldsymbol{T}_1^\top \boldsymbol{D}_*\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1 \\
&\qquad = \boldsymbol{D}_*^{-1/2}\{I_{pJ} - \boldsymbol{D}_*^{1/2}\boldsymbol{T}_1(\boldsymbol{T}_1^\top \boldsymbol{D}_*\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1\boldsymbol{D}_*^{1/2}\}\boldsymbol{D}_*^{-1/2}.
\end{aligned}$$

Then, since $I_{pJ} - \boldsymbol{D}_*^{1/2}\boldsymbol{T}_1(\boldsymbol{T}_1^\top \boldsymbol{D}_*\boldsymbol{T}_1)^{-1}\boldsymbol{T}_1\boldsymbol{D}_*^{1/2}$ is the projection onto the orthogonal complement of the column space of $\boldsymbol{D}_*^{1/2}\boldsymbol{T}_1$, it is positive semidefinite. Thus $\mathrm{avar}[\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{B}} - \boldsymbol{B}_*)\}] - \mathrm{avar}[\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{B}}_{r_*} - \boldsymbol{B}_*)\}]$ is positive semidefinite. The conclusion follows from the fact that for any positive semidefinite matrix $\boldsymbol{Q}$, $\boldsymbol{E}^\top\boldsymbol{Q}\boldsymbol{E} \succeq 0$ for matrix of basis vectors in $\mathbb{R}^{pJ}$, $\boldsymbol{E} \in \mathbb{R}^{J\times pJ}$.

# E Auxiliary lemmas

In this section, we prove two lemmas used in the proof of Theorem 1.

**Lemma E.1.** *Recall $R_{(j)}$ and $\boldsymbol{b}_{(j)}(\cdot)$ defined in* (16) *and* (20)*, respectively. Under assumptions of Theorem 1, we have*

$$\begin{aligned}
&\log R_{(j)}\big(s, \boldsymbol{b}_{(j)}(\boldsymbol{\alpha} + \boldsymbol{\nu})\big) - \log R_{(j)}\big(s, \boldsymbol{b}_{(j)}(\boldsymbol{\alpha})\big) \qquad (36)\\
&\qquad = \bar{\boldsymbol{x}}_{(j)}(s)^\top \boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}) + \frac{1}{2}\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu})^\top \mathbf{F}_{(j)}(s)\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}) + r_{(j)}(\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}), s),
\end{aligned}$$

*where*

$$\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}) := \boldsymbol{b}_{(j)}(\boldsymbol{\alpha} + \boldsymbol{\nu}) - \boldsymbol{b}_{(j)}(\boldsymbol{\alpha}) = (\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\alpha}_{\widetilde{\boldsymbol{U}}} + (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}} + (\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^\top \otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}.$$

*Furthermore,*

$$r_n(\boldsymbol{\nu}) = O_p\left(\frac{\|\boldsymbol{\nu}\|_2^2}{\sqrt{n}} + \frac{\|\boldsymbol{\nu}\|_2^4}{n} + \frac{\|\boldsymbol{\nu}\|_2^3}{\sqrt{n}}\left[1 + \left(\frac{\|\boldsymbol{\nu}\|_2}{\sqrt{n}}\right)^3\right]\right), \tag{37}$$

*where $r_n(\cdot)$ is defined in* (25)*.*

*Proof.* Recall that for every $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$R_{(j)}(s, \boldsymbol{\beta}) = \sum_{i=1}^{n_{(j)}} \mathbf{1}(y_{(j)i} \geq s) \exp\left(\boldsymbol{x}_{(j)i}^\top \boldsymbol{\beta}\right).$$

The main expansion in (36) follows directly from differentiation and a Taylor series expansion. We will now prove (37).

Lemma A2 of Hjort and Pollard (2011) allows us an expansion of $\log R_{(j)}\big(s, \boldsymbol{b}_{(j)}(\boldsymbol{\alpha} + \boldsymbol{\nu})\big) = \log R_{(j)}\big(s, \boldsymbol{b}_{(j)}(\boldsymbol{\alpha}) + \boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu})\big)$ around $\log R_{(j)}\big(s, \boldsymbol{b}_{(j)}(\boldsymbol{\alpha})\big)$. For each $j \in [J]$, using $w_i = \mathbf{1}(y_{(j)i} \geq s) \exp(\boldsymbol{x}_{(j)i}^\top \boldsymbol{b}_{(j)}(\boldsymbol{\alpha}))$, $a_i = \boldsymbol{x}_{(j)i}^\top \boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu})$, and $t = 1$ in Lemma A2 of Hjort and Pollard (2011), for all $j \in [J]$, we get (36) where

$$\begin{aligned}
|r_{(j)}(\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}), s)| &\leq \frac{4}{3} \max_{i \leq n_{(j)}} |(\boldsymbol{x}_{(j)i} - \bar{\boldsymbol{x}}_{(j)}(s))^\top \boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu})|^3 \\
&\leq \frac{4}{3}(2K)^3 \|\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu})\|_2^3 \\
&\leq \frac{4}{3} J (2K)^3 (2\|\boldsymbol{\alpha}\|_2^3 \|\boldsymbol{\nu}\|_2^3 + \|\boldsymbol{\nu}\|_2^6),
\end{aligned}$$

where $K$ is the absolute bound on the predictors and the last inequality follows by an application of Cauchy-Schwarz inequality on (23). Thus

$$\begin{aligned}
&\sum_{j=1}^{J} \sum_{i=1}^{n_{(j)}} \int_0^L r_{(j)}\big(\boldsymbol{\gamma}_{(j)}(\boldsymbol{\alpha}, \boldsymbol{\nu}/\sqrt{n}), s\big) dN_{(j)i}(s) \\
&\quad \leq \frac{4}{3} J (2K)^3 \left(2\|\boldsymbol{\alpha}\|_2 \frac{\|\boldsymbol{\nu}\|_2^3}{n^{3/2}} + \frac{\|\boldsymbol{\nu}\|_2^6}{n^3}\right) \sum_{j=1}^{J} \sum_{i=1}^{n_{(j)}} dN_{(j)i}(s) \\
&\quad \leq \frac{4}{3} J (2K)^3 \left(2\|\boldsymbol{\alpha}\|_2 \frac{\|\boldsymbol{\nu}\|_2^3}{n^{1/2}} + \frac{\|\boldsymbol{\nu}\|_2^6}{n^2}\right) \\
&\quad = O_p\left(\frac{\|\boldsymbol{\nu}\|_2^3}{n^{1/2}}\Big(1 + \left[\frac{\|\boldsymbol{\nu}\|_2}{\sqrt{n}}\right]^3\Big)\right).
\end{aligned}$$

In Lemma E.2, we show that $\boldsymbol{A} = (\boldsymbol{a}_{(1)}^\top, \boldsymbol{a}_{(2)}^\top, \cdots, \boldsymbol{a}_{(J)}^\top)^\top = O_p(1)$ and $\boldsymbol{D}_{(j)} = O_p(1)$. Thus (37)

follows by observing

$$n^{-1/2}\sum_{j=1}^{J}\boldsymbol{a}_{(j)}^{\top}(\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^{\top}\otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}} = O_p(n^{-1/2}\|\boldsymbol{\nu}\|_2^2),$$

$$\frac{1}{2n}\sum_{j=1}^{J}\big((\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^{\top}\otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}\big)^{\top}\boldsymbol{D}_{(j)}\big((\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^{\top}\otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}\big) = O_p(n^{-1}\|\boldsymbol{\nu}\|_2^4),$$

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{J}\big((\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^{\top}\otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}\big)^{\top}\boldsymbol{D}_{(j)}\big((\boldsymbol{\nu}_{\widetilde{\boldsymbol{V}}_{(j)}}^{\top}\otimes I_p)\boldsymbol{\alpha}_{\widetilde{\boldsymbol{U}}} + (\boldsymbol{\alpha}_{\widetilde{\boldsymbol{V}}_{(j)}}^{\top}\otimes I_p)\boldsymbol{\nu}_{\widetilde{\boldsymbol{U}}}\big) = O_p(n^{-1/2}\|\boldsymbol{\nu}\|_2^3). \quad \square$$

**Lemma E.2.** *Under assumptions **(A2)** – **(A7)**,*

$$\boldsymbol{D}_{(j)} \xrightarrow{p} \kappa_{(j)}\boldsymbol{D}_{*(j)} \qquad \textit{and} \qquad \boldsymbol{A} \xrightarrow{d} \mathrm{N}(0, \boldsymbol{D}_*),$$

*where $\boldsymbol{A}$ and $\boldsymbol{D}$ are defined in* (27) *and* (28)*, respectively, and $\boldsymbol{D}_{*(j)} := \int_0^L \boldsymbol{D}_{*(j)}(s)dH_{(j)}(s)$, $\boldsymbol{D}_* := \mathrm{BlockDiag}\{\kappa_{(j)}\boldsymbol{D}_{*(j)}\}_{j=1}^{J}$, with $\kappa_{(j)}$ is defined in **(A5)**, and $\boldsymbol{D}_{*(j)}$ is the population information matrix for the $j$th population (when analyzed independently of the other $J-1$ populations), i.e., $\widetilde{\boldsymbol{b}}_{(j)}$, the unconstrained MLE for $\boldsymbol{b}_{*(j)}$, has the following property*

$$\sqrt{n_{(j)}}(\widetilde{\boldsymbol{b}}_{(j)} - \boldsymbol{b}_{*(j)}) \xrightarrow{d} \mathrm{N}\big(0, \boldsymbol{D}_{*(j)}^{-1}\big).$$

*Proof.* Recall that $\boldsymbol{A} = [\boldsymbol{a}_{(1)}^{\top}, \boldsymbol{a}_{(2)}^{\top}, \cdots, \boldsymbol{a}_{(J)}^{\top}]^{\top}$, where $\boldsymbol{a}_{(j)} = n^{-1/2}\sum_{i=1}^{n_{(j)}}\int_0^L(\boldsymbol{x}_{(j)i} - \bar{\boldsymbol{x}}_{(j)}(s))dN_{(j)i}(s) \in \mathbb{R}^p$, see (24). By assumption **(A7)** and the proof of (i) in Theorem 6.1 of Hjort and Pollard (2011), we have that $n\boldsymbol{D}_{(j)}/n_{(j)} \xrightarrow{p} \boldsymbol{D}_{*(j)}$. Define $\boldsymbol{C}_{(j)} := \{n/n_{(j)}\}^{1/2}\boldsymbol{a}_{(j)}$. Then by proof of Theorem 6.1 of Hjort and Pollard (2011) (see (6.8)–(6.10)), $\boldsymbol{C}_{(j)} \xrightarrow{d} \mathrm{N}(0, \boldsymbol{D}_{*(j)})$, thus $\boldsymbol{a}_{(j)} \xrightarrow{d} \mathrm{N}(0, \kappa_{(j)}\boldsymbol{D}_{*(j)})$, by Slutsky's theorem. Furthermore, as the $J$ populations are independent by **(A6)**, we have that $\mathrm{cor}(\boldsymbol{a}_{(j)}, \boldsymbol{a}_{(j')}) = 0$, from which the result follows. $\square$

# F Additional simulation studies

## F.1 Weighted partial likelihood estimator

At the suggestion of a referee, we considered an alternative estimator based on a sample-size weighted partial log-likelihood. In particular, we also considered the estimator

$$\underset{\boldsymbol{B}\in\mathcal{C}_r\cap\mathcal{A}_s}{\arg\min}\big\{-\boldsymbol{\mathcal{L}}^{\mathrm{W}}(\boldsymbol{B}) + \mu\|\boldsymbol{B}\|_F^2\big\} \tag{38}$$

where $\mathcal{C}_r$, $\mathcal{A}_s$, and $\mu$ are as defined in the main manuscript, and

$$\boldsymbol{\mathcal{L}}^{\mathrm{W}}(\boldsymbol{B}) = \sum_{j=1}^{J}\frac{1}{n_{(j)}}\sum_{i=1}^{n_{(j)}}\delta_{(j)i}\left[\boldsymbol{x}_{(j)i}^{\top}\boldsymbol{b}_{(j)} - \log\big\{\sum_{k\in\mathcal{R}_{(j)i}}\exp(\boldsymbol{x}_{(j)i}^{\top}\boldsymbol{b}_{(j)})\big\}\right]$$

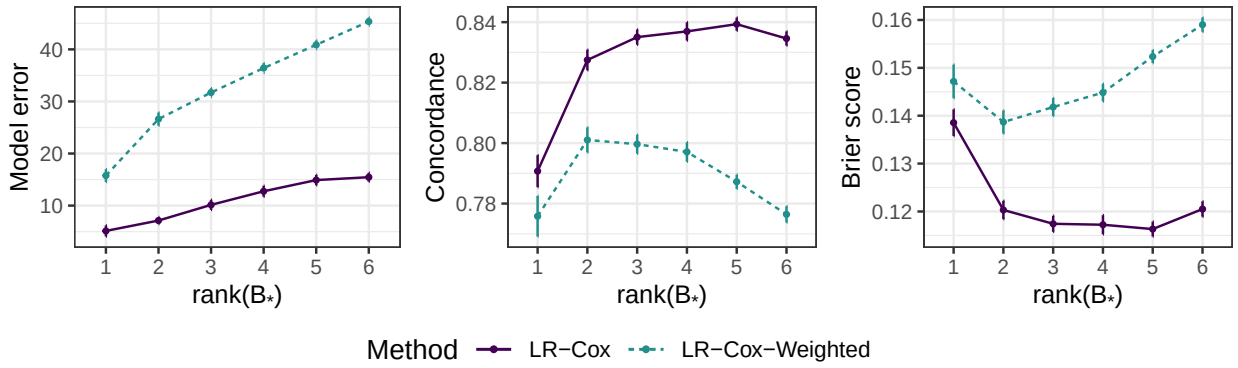


Figure 1: Comparison of (3) versus (38) for estimating the entire coefficient matrix $\boldsymbol{B}_*$ under the data generating model described in Section 5.1 of the main manuscript.

is the weighted partial log-likelihood. The estimator (38), in contrast to (3), scales each populations' contribution to the log-likelihood by its sample size so that each population contributes to the partial log-likelihood in an approximately equivalent manner. In this section, we explore how (38), which we call `LR-Cox-Weighted`, compares to (3) under the same data generating models considered in Section 5 of the main manuscript.

Results are displayed in Figures 1 and 2. In Figure 1, we see that (38) performs substantially worse than (3) in all three metrics. As rank($\boldsymbol{B}_*$) increases, (38) begins to perform worse than `Convex-Approx` (not included in Figure 1, but can be seen in the main manuscript). In Figure 2, we compare model errors for the coefficient vectors $\boldsymbol{b}_{*(J-2)}, \boldsymbol{b}_{*(J-1)}$, and $\boldsymbol{b}_{*(J)}$ which correspond to populations with sample sizes 100, 200, and 300, respectively. In Figure 2, we see that sample sizes tend to affect errors more substantially for the estimator (38) than they do for (3). For example, the model error of (3) in population $(J-2)$ is more than double that in population $J$. This is not the case for (38): errors are nearly equivalent across all populations. However, this is a moot point since in each population, (3) significantly outperforms (38).

## F.2 Sensitivity to the choice of rank

In both our simulation studies and real data application, the rank parameter, $r$, along with the number of predictors to include in the model, $s$, were both chosen by cross-validation. A referee suggested we examine consider how (3) performs if one overspecifies the rank and tunes only $s$. We did exactly this under the same simulation settings as those in the top panel of Figure 1 of the main manuscript. Specifically, in Figure 3 we display results for our method with $(r, s)$ chosen via a validation set (`LR-Cox-CV`); and our method with $s$ chosen using the validation set and $r \in \{1, 2, 3, 5, 7\}$ fixed (`LR-Cox-1`, `LR-Cox-2`, `LR-Cox-3`, `LR-Cox-5`, `LR-Cox-7`, respectively). Interestingly, we see that the tuned version of our method tends to perform similarly to the version of our method with correctly specified rank. It is notable that the version of our method with $r = 7$ – which is overspecified in every setting considered – does not perform much worse than our method with $r$ chosen to minimize the validation likelihood. This would suggest that if one has a reasonable sense of the rank of $\boldsymbol{B}_*$, slightly

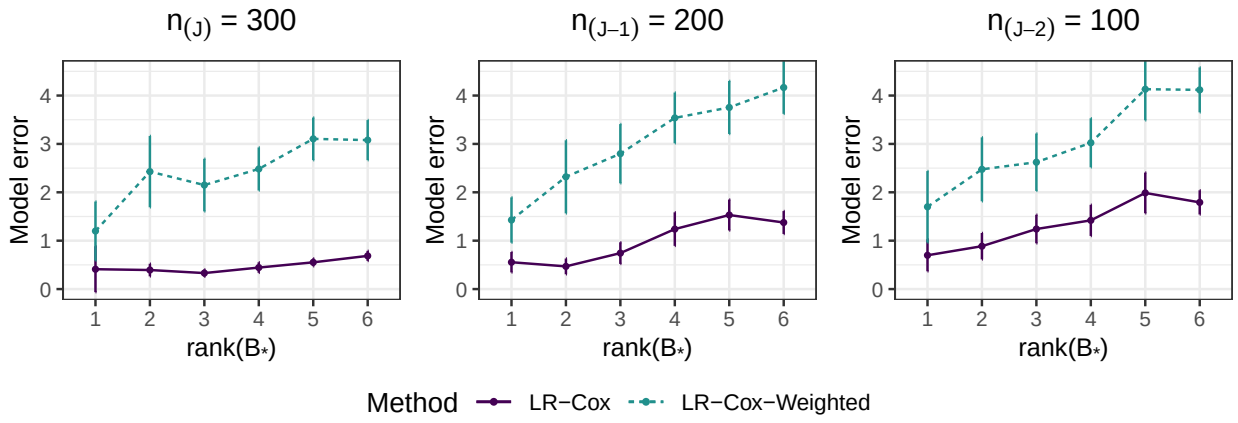


Figure 2: Comparison of (3) versus (38) for estimating estimating (left) $\boldsymbol{b}_{*(J)}$, (center) $\boldsymbol{b}_{*(J-1)}$, and (right) $\boldsymbol{b}_{*(J-2)}$ under the data generating model described in Section 5.1 of the main manuscript. For $\boldsymbol{b}_{*(J)}$, model error is defined as $\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{b}_{*(J)} - \widehat{\boldsymbol{b}}_{(J)})\|_2^2$, and similarly for $\boldsymbol{b}_{*(J-1)}$ and $\boldsymbol{b}_{*(J-2)}$

overspecifying $r$ and tuning only $s$ may be a reasonable approach if computing time is an issue.

## F.3 Alternative data generating models

In this subsection, we perform additional simulation studies to analyze the performance of our method when there are varying degrees of cancer-specific and shared factors. In each of the studies, we generate data from a model in which some factors are relevant to only a subset of populations (e.g., cancer types). Throughout this section, for a set $\mathcal{S} \subset [p]$, we use $\boldsymbol{C}_{\mathcal{S},\cdot}$ (resp. $\boldsymbol{c}_{\mathcal{S}}$) to denote to the submatrix (resp. subvector) of $\boldsymbol{C}$ (resp. $\boldsymbol{c}$) consisting only of the rows (resp. elements) indexed by $\mathcal{S}$. In each of the following models, Model A–C, we set $r_* = 6$. As in the main paper and in the earlier simulations considered in this Web Appendix, for each of the model, we simulate one hundred independent replications, we generate survival times under the Cox proportional hazards models for $J = 12$ distinct populations. In each setting, we generate $n_{(1)} = n_{(4)} = n_{(7)} = n_{(10)} = 100$, $n_{(2)} = n_{(5)} = n_{(8)} = n_{(11)} = 200$, and $n_{(3)} = n_{(6)} = n_{(9)} = n_{(12)} = 300$ independent failure times for each population. Here, we fix $\mu = 1$ and $\rho_0 = 50$ for our method, and use all other methods as they are described in the main manuscript.

The first model we consider is Model A.

- **Model A** (Partially shared factors, distinct predictors): We partition the $J = 12$ populations into two groups of size $J-q$ and $q$ for $q \in \{1, \ldots, 5\}$. We randomly select 20 of the $p$ predictors to be relevant: 10 predictors are relevant for the first group and second group, respectively, and these sets are mutually exclusive. Let these sets of predictors be denoted $\mathcal{S}_1$ and $\mathcal{S}_2$. Then, we set $(\boldsymbol{b}_{*(1)}, \ldots, \boldsymbol{b}_{*(J-q)})_{\mathcal{S}_1,\cdot} = \{\sqrt{(J-q)/2}\} \cdot \boldsymbol{U}_1 \boldsymbol{V}_1^\top$ where $\boldsymbol{U}_1 \in \mathbb{R}^{10 \times (r_*-1)}$ has iid entries from Uniform$(1, 2)$ and $\boldsymbol{V}_1 \in \mathbb{R}^{(J-q)\times(r_*-1)}$ is a randomly generated semiorthogonal matrix such that $\boldsymbol{V}_1^\top \boldsymbol{V}_1 = I_{r_*-1}$. We set all other

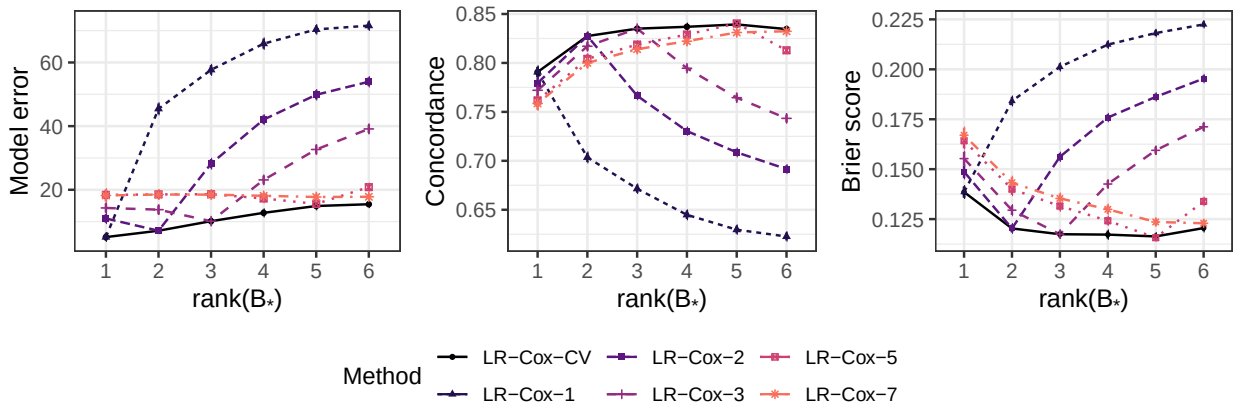


Figure 3: Model error, concordance, and Brier score for our method with $(r, s)$ chosen via a validation set (`LR-Cox-CV`); and our method with $s$ chosen using the validation set and $r \in \{1, 2, 3, 5, 7\}$ fixed (`LR-Cox-1`, `LR-Cox-2`, `LR-Cox-3`, `LR-Cox-5`, `LR-Cox-7`, respectively).

entries of $(\boldsymbol{b}_{*(1)}, \ldots, \boldsymbol{b}_{*(J-q)})$ equal to zero. Separately, we set $(\boldsymbol{b}_{*(J-q+1)}, \ldots, \boldsymbol{b}_{*(12)})_{\mathcal{S}_2,\cdot} = (\sqrt{q/2}) \cdot \boldsymbol{U}_2 \boldsymbol{V}_2^\top$ where $\boldsymbol{U}_2 \in \mathbb{R}^{10\times 1}$ has iid entries from Uniform$(1, 2)$ and $\boldsymbol{V}_2 \in \mathbb{R}^{q\times 1}$ is uniformly distributed on the unit sphere. All other entries of $(\boldsymbol{b}_{*(J-q+1)}, \ldots, \boldsymbol{b}_{*(12)})$ are set equal to zero. The matrices $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are generated entirely independently, as are $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$.

Under Model A, $J - q$ populations share $r_* - 1$ common factors and $q$ populations share one common factor, but no factors are shared between the two groups. Also, the factors for each of the two groups depend on entirely distinct sets of proteins. We consider $q \in \{1, \ldots, 5\}$ with $p = 500$ and $p \in \{100, 200, \ldots, 500\}$ with $q = 3$.

We present results for 100 independent replications under Model A in Figures 4 and 5. In Figure 4 we see that with $p$ varying and $q = 3$, and with $q$ varying and $p = 500$, our method substantially outperforms competitors in terms of model error, concordance, and Brier score across the $J$ populations. In Figure 5, we examine how each method estimates individual columns of $\boldsymbol{B}_*$, focusing on $\boldsymbol{b}_{*(J)}, \boldsymbol{b}_{*(J-1)}$, and $\boldsymbol{b}_{*(J-2)}$. We see that even though these populations share only a single factor with one another, our method outperforms all of the competing methods.

Next, we consider another data generating model, Model B.

– **Model B** (Partially shared factors, partially shared predictors): We partition the $J = 12$ populations into two groups of size eight and four. We randomly select 15 of the $p$ predictors to be relevant. Then, we randomly allocate these 15 predictors into two sets of ten elements each, $\mathcal{S}_1$ and $\mathcal{S}_2$, such that cardinality of their intersection is five. With $q \in \{1, \ldots, 5\}$, we set $(\boldsymbol{b}_{*(1)}, \ldots, \boldsymbol{b}_{*(8)})_{\mathcal{S}_1,\cdot} = (\sqrt{(r_* - q)/2}) \cdot \boldsymbol{U}_1 \boldsymbol{V}_1^\top$ where $\boldsymbol{U}_1 \in \mathbb{R}^{10\times(r_*-q)}$ and $\boldsymbol{V}_1 \in \mathbb{R}^{8\times(r_*-q)}$ in the same manner as in Model A. All other entries of $(\boldsymbol{b}_{*(1)}, \ldots, \boldsymbol{b}_{*(8)})$ are set to zero. Separately, we set $(\boldsymbol{b}_{*(9)}, \ldots, \boldsymbol{b}_{*(12)})_{\mathcal{S}_2,\cdot} = (\sqrt{q/2}) \cdot \boldsymbol{U}_2 \boldsymbol{V}_2^\top$ where $\boldsymbol{U}_2 \in \mathbb{R}^{10\times q}$ and $\boldsymbol{V}_2 \in \mathbb{R}^{4\times q}$ are generated as in Model A. The matrices $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are generated entirely independently, as are $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$.

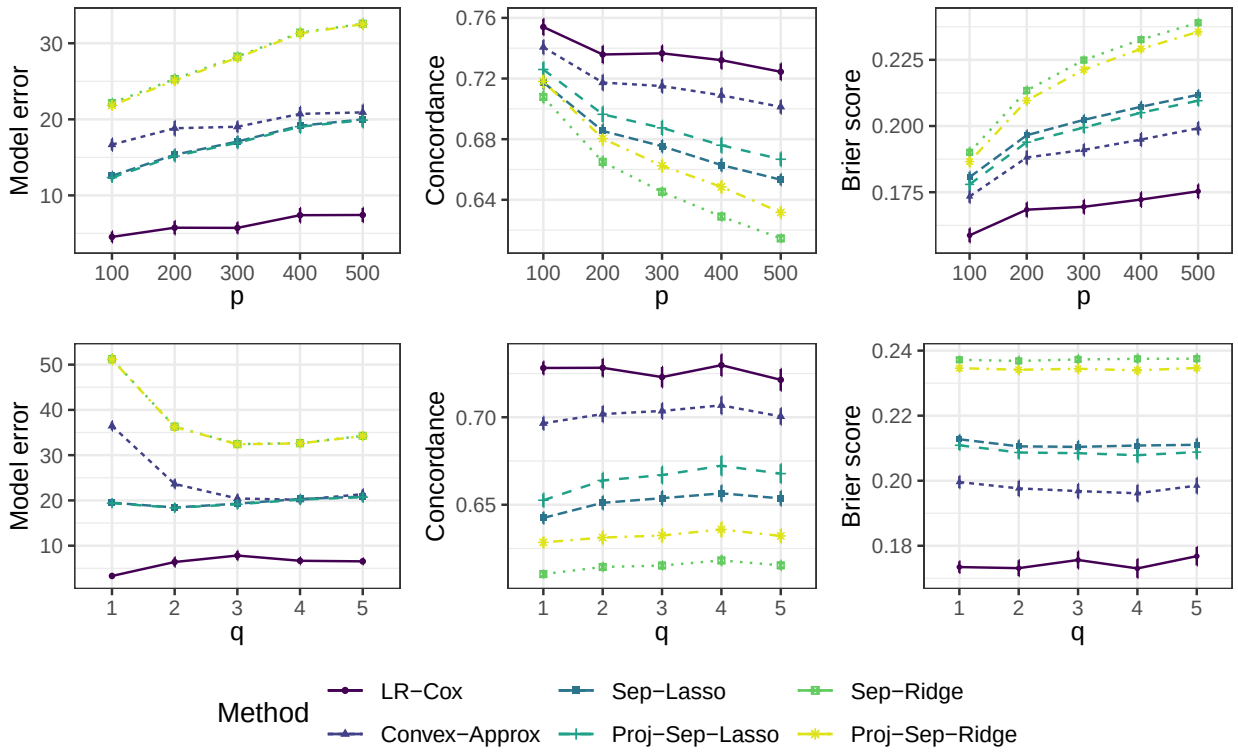


Figure 4: Average (plus/minus two standard errors) model error, concordance, and Brier score with (top row) $(p, q) \in \{100, 200, \ldots, 500\} \times \{3\}$ and (bottom row) $(p, q) \in \{500\} \times \{1, \ldots, 5\}$ under Model A.

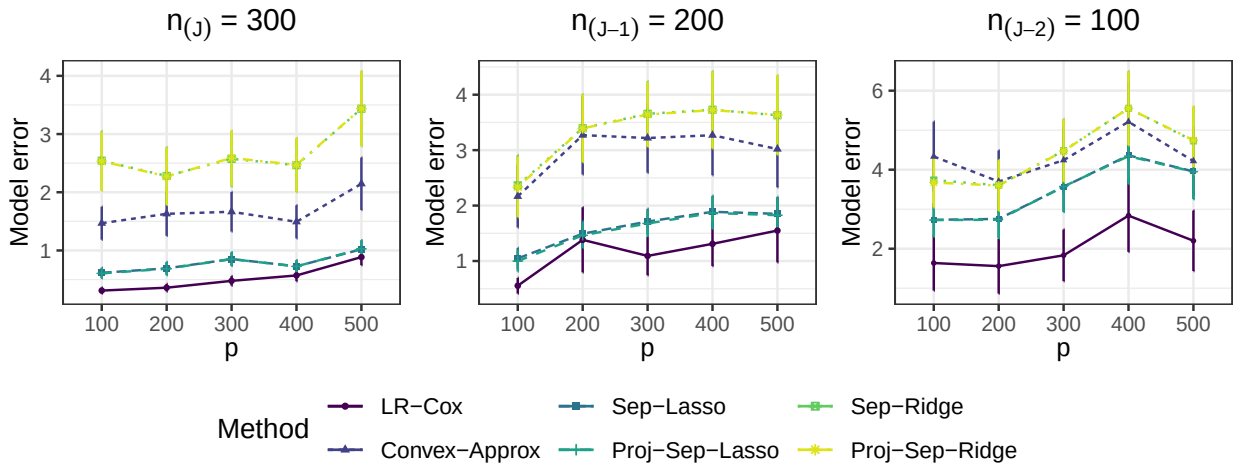


Figure 5: Average model error (plus/minus two standard errors) for (left) $\boldsymbol{b}_{*(J)}$, (center) $\boldsymbol{b}_{*(J-1)}$, and (right) $\boldsymbol{b}_{*(J-2)}$ with $p \in \{100, 200, \ldots, 500\}$ and $q = 3$ under Model A. For $\boldsymbol{b}_{*(J)}$, model error is defined as $\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{b}_{*(J)} - \widehat{\boldsymbol{b}}_{(J)})\|_2^2$, and similarly for $\boldsymbol{b}_{*(J-1)}$ and $\boldsymbol{b}_{*(J-2)}$.

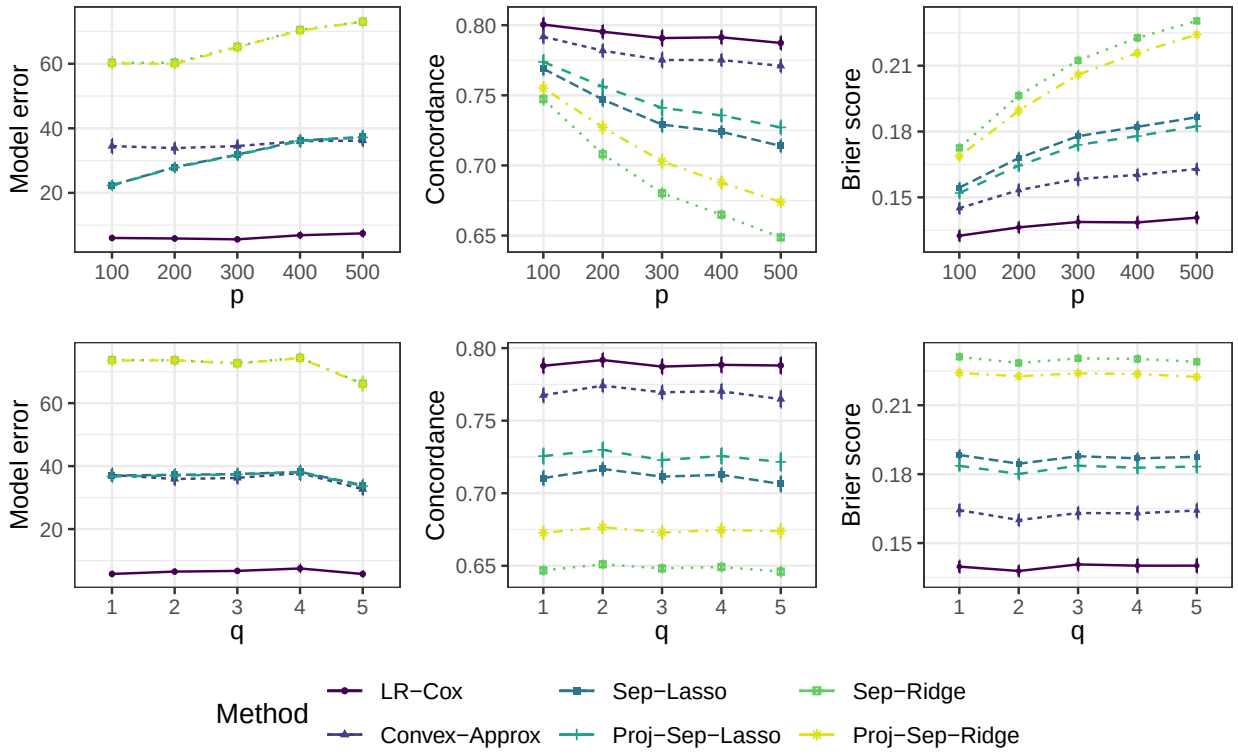


Figure 6: Average (plus/minus two standard errors) model error, concordance, and Brier score with (top row) $(p, q) \in \{100, 200, \ldots, 500\} \times \{3\}$ and (bottom row) $(p, q) \in \{500\} \times \{1, \ldots, 5\}$ under Model B.

Under Model B, eight populations share $r_* - q$ common factors and four populations share $q$ common factors, but no factors are shared between the two groups. Important predictors are partially shared, but each of the two groups of populations have five important predictors which are irrelevant for the other group.

Results based on 100 independent replications are displayed in Figures 6 and 7. In Figure 6, we see that `LR-Cox` outperforms all competitors in all scenarios considered. When examining the results in Figure 7, we see that `Sep-Lasso` performs nearly as well as `LR-Cox` for estimating the $J$th population's regression coefficients (in terms of model error), but for the smaller sample size populations (e.g., the $(J-1)$th and $(J-2)$th) our method more substantially outperforms `Sep-Lasso`.

Finally, we consider Model C.

– **Model C** (Distinct factors, partially shared predictors): We partition the $J = 12$ populations into two groups of size $J - t$ and $t$ for $t \in \{0, 1, \ldots, 4\}$. We randomly construct sets $\mathcal{S}_l$ for $l \in \{1, \ldots, t+1\}$ so that $\mathcal{S}_1$ consists of ten randomly chosen elements of $[p]$, and $\mathcal{S}_l$ for $l > 1$ consists of five randomly chosen elements of $\mathcal{S}_1$ and five randomly chosen elements of $[p] \setminus \mathcal{S}_1$. Then, we set $(\boldsymbol{b}_{*(1)}, \ldots, \boldsymbol{b}_{*(J-t)})_{\mathcal{S}_1,\cdot} = (\sqrt{(J-t)/2}) \cdot \boldsymbol{U}_1 \boldsymbol{V}_1^\top$ where $\boldsymbol{U}_1 \in \mathbb{R}^{10 \times (r_* - t)}$ and $\boldsymbol{V}_1 \in \mathbb{R}^{(J-t) \times (r_* - t)}$ in the same manner as in Model A. All other entries of $(\boldsymbol{b}_{*(1)}, \ldots, \boldsymbol{b}_{*(J-t)})$ are set to zero. Finally, letting $\check{l} = l - (J - t + 2)$, for $l \in \{J - t + 1, \ldots, J\}$, we set $\boldsymbol{b}_{*(l)\mathcal{S}_{\check{l}}}$ to have entries which are iid

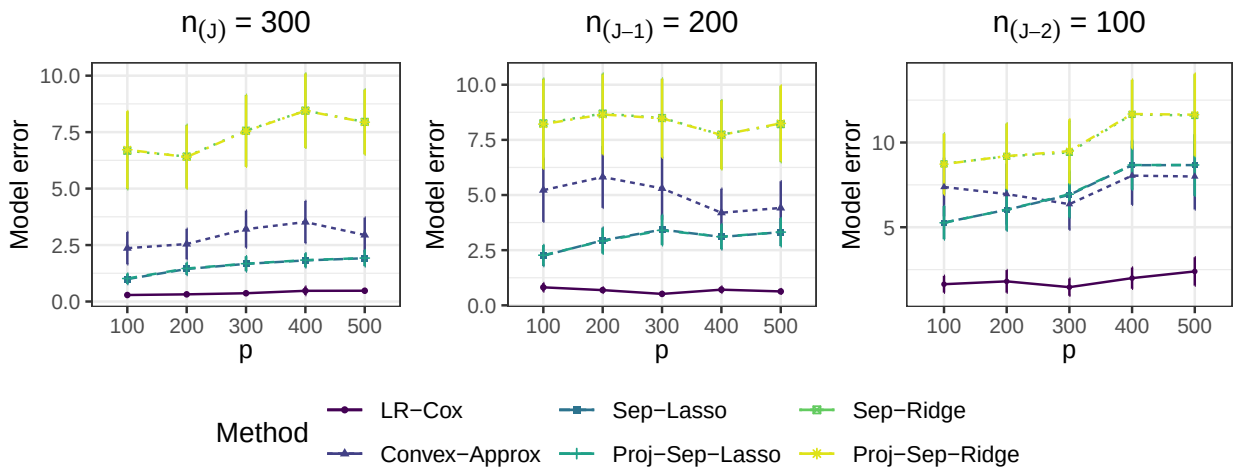


Figure 7: Average (plus/minus two standard errors) model error for (left) $\boldsymbol{b}_{*(J)}$, (center) $\boldsymbol{b}_{*(J-1)}$, and (right) $\boldsymbol{b}_{*(J-2)}$ with $p \in \{100, 200, \ldots, 500\}$ and $q = 3$ under Model B. For $\boldsymbol{b}_{*(J)}$, model error is defined as $\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{b}_{*(J)} - \widehat{\boldsymbol{b}}_{(J)})\|_2^2$, and similarly for $\boldsymbol{b}_{*(J-1)}$ and $\boldsymbol{b}_{*(J-2)}$.

$\text{Uniform}\{[-2\sqrt{2}, -\sqrt{2}] \cup [\sqrt{2}, 2\sqrt{2}]\}$. All other entries of the $\boldsymbol{b}_{*(l)}$ are set to zero.

Model C is essentially the worst case scenario for our method. In particular, $t$ populations do not share any factors with the other $J-1$ populations. Moreover, each of these $t$ populations' factors depend on a partially distinct set of predictors.

We display results for 100 independent replications under Model C in Figures 8 and 9. In Figure 8, we see that when $t$ is relatively small, our method outperforms the competitors. When $t$ is larger (e.g., $t \geq 3$), we see that `LR-Cox` can be outperformed by `Sep-Lasso`, `Proj-Sep-Lasso`, and `Convex-Approx`. In terms of estimating the population-specific regression coefficients, we see that `LR-Cox` is significantly worse than `Sep-Lasso` and `Proj-Sep-Lasso` in both populations with $n_{(j)} \leq 200$. This is not surprising given that these populations do not have any factors in common with the other $J-1$ populations, so our method may impose unhelpful bias relative to the methods which fit a model for each population separately. In these extreme cases, it may be helpful to also tune $\mu$, since evidently both `Sep-Lasso` and `Proj-Sep-Lasso` benefit from shrinkage.

## F.4 Additional performance metrics

In this subsection, we provide additional results from the simulation studies detailed in Section 5 of the main manuscript. Specifically, in Figure 10, we display results using mean squared error, $\|\widehat{\boldsymbol{B}} - \boldsymbol{B}_*\|_F^2/(pJ)$, as a performance metric for each of the six methods under the different scenarios represented by each row of Figure 1 of the main manuscript. In Figure 10 we see that with $p$ fixed, the relative performances in terms of mean squared error mostly agree with those using model error as the performance metric. The main difference comes in the right-most panel of Figure 10, where mean squared error decreases as $p$ increases. This is because model error (as we define it), $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}_*)\|_F^2$, does not adjust for the dimensionality $p$ as does mean squared error $\|\widehat{\boldsymbol{B}} - \boldsymbol{B}_*\|_F^2/(pJ)$.

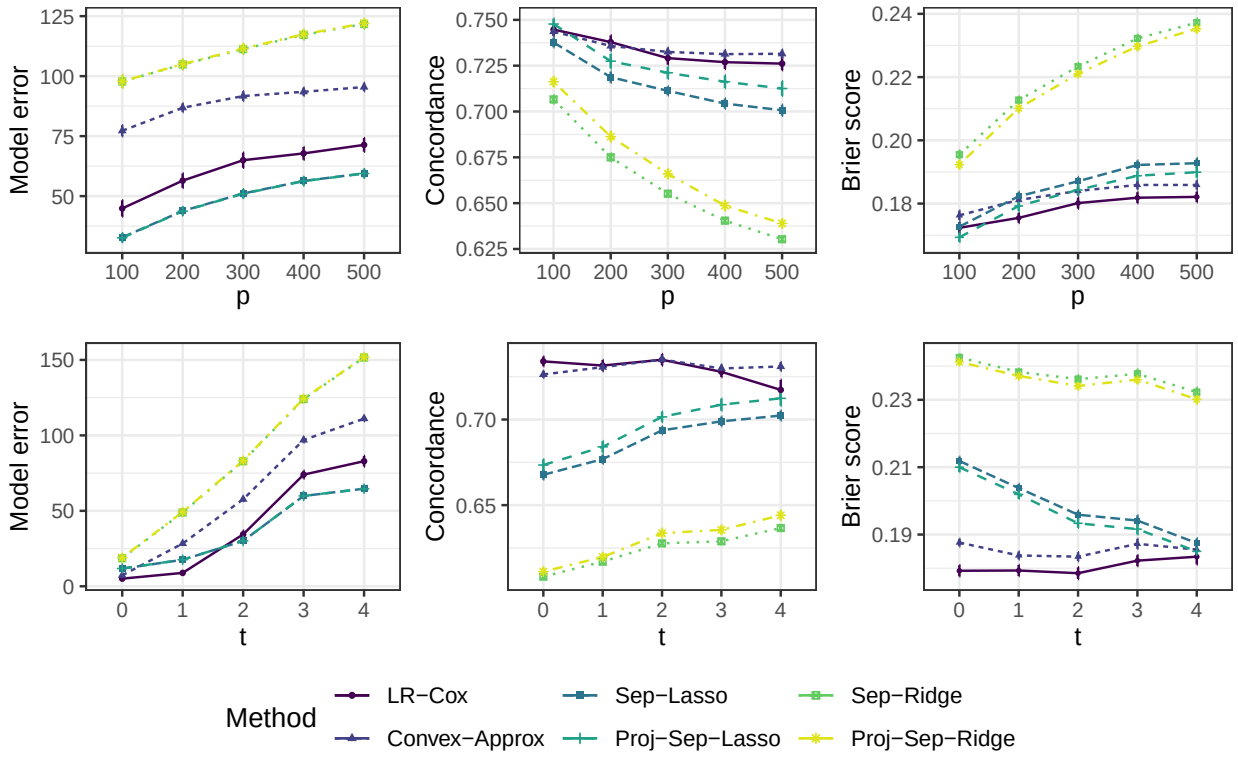


Figure 8: Average (plus/minus two standard errors) model error, concordance, and Brier score with (top row) $(p, t) \in \{100, 200, \ldots, 500\} \times \{3\}$ and (bottom row) $(p, t) \in \{500\} \times \{0, \ldots, 4\}$ under Model C.

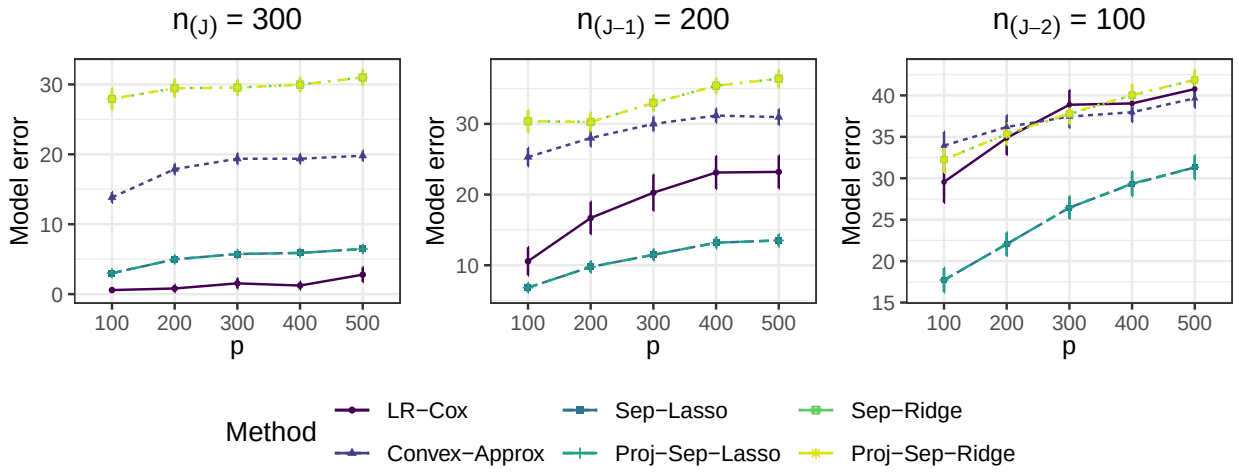


Figure 9: Average (plus/minus two standard errors) model error for (left) $\boldsymbol{b}_{*(J)}$, (center) $\boldsymbol{b}_{*(J-1)}$, and (right) $\boldsymbol{b}_{*(J-2)}$ with $p \in \{100, 200, \ldots, 500\}$ and $t = 3$ under Model C. For $\boldsymbol{b}_{*(J)}$, model error is defined as $\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{b}_{*(J)} - \widehat{\boldsymbol{b}}_{(J)})\|_2^2$, and similarly for $\boldsymbol{b}_{*(J-1)}$ and $\boldsymbol{b}_{*(J-2)}$.

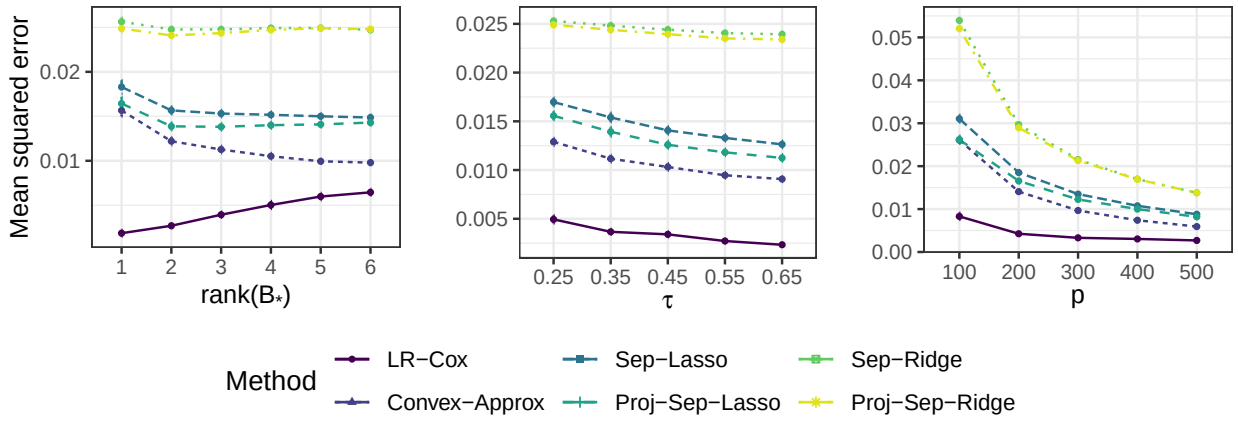


Figure 10: Average (plus/minus two standard errors) mean squared error $\|\boldsymbol{B}_* - \widehat{\boldsymbol{B}}\|_F^2/(pJ)$ with (left) $p = 250$ and $\tau = 0.35$; (center) $p = 250$ and $r_* = 3$; and (right) $r_* = 3$ and $\tau = 0.35$ under the data generating model from Section 5.1 of the main manuscript.

# G Additional real data analysis results

In this section, we perform additional analyses to assess whether our method leads to improved estimation accuracy on each of the five most rare cancer types analyzed in Section 6 of the main manuscript. These are PAAD, ESCA, LIHC, CESC, and GBM, whose sample sizes are 105, 126, 184, 171, and 205, respectively.

For each of the cancer types of interest, we performed leave-one-out cross-validation. Specifically, for each $i \in [n_{(j)}]$ separately, we fit the model using all the data (from all cancer types) except the $i$th subject from the $j$th type. Tuning parameters were chosen by five-fold cross-validation on the training data. Then, with the fitted model, we obtained an estimate of $\boldsymbol{x}_{(j)i}^\top \boldsymbol{b}_{*(j)}$, which we call $\widehat{\phi}_{(j)i}$. Once we have done this for each of the $n_{(j)}$ subjects, we compute (a) the concordance between the estimated linear predictors $\{\widehat{\phi}_{(j)1}, \ldots, \widehat{\phi}_{(j)n_{(j)}}\}$ and the true event times in the $j$th dataset; and (b) the linear predictor score, which we define as

$$-2 \sum_{i=1}^{n_{(j)}} \delta_{(j)i} \log \left\{ \frac{\exp(\widehat{\phi}_{(j)i})}{\sum_{k \in \mathcal{R}_{(j)i}} \exp(\widehat{\phi}_{(j)k})} \right\},$$

e.g., see Dai and Breheny (2019). To be clear, each $\widehat{\phi}_{(j)i}$ is computed based on a model fit to data which excluded the $i$th subject with the $j$th cancer type. Of course, lower linear predictor score would suggest a better model fit.

We did this using our method, `LR-Cox`, and the three competitors: `Sep-Ridge`, `Sep-Lasso`, and `Sep-En`. Note that the three competitors do not use any information from the other cancer types when fitting the model. Results are displayed in Table 1. Here, we see that `LR-Cox` outperforms the three competitors in terms of concordance in four of the five datasets. However in one dataset, ESCA, none of the methods have concordance above 0.5, which corresponds to randomly guessing the order of the linear predictors. Similarly, `LR-Cox` outperforms competitors in the same four of five datasets in terms of linear predictor score.

Table 1: Leave-one-out cross-validated concordance and linear predictor scores for the four considered methods across the five considered datasets. Bold cells are those with highest concordance or lowest linear predictor score.

| | Concordance | | | | | Linear predictor score | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PAAD | ESCA | LIHC | CESC | GBM | PAAD | ESCA | LIHC | CESC | GBM |
| `LR-Cox` | **0.675** | 0.405 | **0.689** | **0.704** | **0.635** | **206.148** | 268.592 | **496.469** | **432.391** | **560.993** |
| `Sep-Ridge` | 0.560 | 0.492 | 0.574 | 0.613 | 0.609 | 211.489 | 264.459 | 510.781 | 444.691 | 564.757 |
| `Sep-Lasso` | 0.575 | 0.500 | 0.586 | 0.461 | 0.571 | 219.049 | 263.538 | 507.540 | 455.385 | 569.579 |
| `Sep-En` | 0.601 | **0.500** | 0.556 | 0.616 | 0.601 | 210.677 | **263.260** | 509.164 | 440.337 | 564.893 |

Taken together, these results provide strong evidence that our method can yield positive findings, even on the rarest cancer types included in our real data analysis.

# H Comparison to sparse reduced rank regression

The method of Chen and Huang (2012) assumes the data $\{(\boldsymbol{y}_i, \boldsymbol{x}_i)\}_{i=1}^n$ consist of a $q$-dimensional (continuous) response variable $\boldsymbol{y}_i$ and $p$-dimensional predictor $\boldsymbol{x}_i$ for $i \in [n]$. The goal is to estimate $\boldsymbol{\beta}_*$ from the model which assumes $\boldsymbol{y}_i$ is a realization of the random vector

$$\boldsymbol{\beta}_*^\top \boldsymbol{x}_i + \epsilon_i, \quad \epsilon_i \in \mathbb{R}^q, \quad \mathrm{E}(\epsilon_i) = 0, \quad \mathrm{Cov}(\epsilon_i) \in \mathbb{S}_+^q,$$

for $i \in [n]$ and assumes that $\boldsymbol{\beta}_* = \boldsymbol{U}\boldsymbol{C}$ for $\boldsymbol{C} \in \mathbb{R}^{r_* \times q}$ and $\boldsymbol{U} \in \mathbb{R}^{p \times r_*}$ where $r_* < \min(p, q)$. Sparse reduced-rank regression, as proposed in Chen and Huang (2012), assumes $\boldsymbol{U}$ is row-wise sparse. Letting $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n \times p}$, $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^\top \in \mathbb{R}^{n \times q}$, and $\|\boldsymbol{U}\|_{1,2} = \sum_{j=1}^p \|\boldsymbol{U}_{j,\cdot}\|_2$, the estimator of $(\boldsymbol{U}, \boldsymbol{C})$ proposed by Chen and Huang (2012) is

$$\underset{\boldsymbol{U} \in \mathbb{R}^{p \times r}, \boldsymbol{C} \in \mathbb{R}^{r \times q}}{\arg\min} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{U}\boldsymbol{C}\|_F^2 + \gamma \|\boldsymbol{U}\|_{1,2} \right\} \text{ subject to } \boldsymbol{C}\boldsymbol{C}^\top = I_r,$$

or equivalently, letting $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q) \in \mathbb{R}^{r \times q}$ and $\boldsymbol{Y}_{\cdot,j} \in \mathbb{R}^n$ denote the $j$th column of $\boldsymbol{Y}$,

$$\underset{\boldsymbol{U} \in \mathbb{R}^{p \times r}, \boldsymbol{C} \in \mathbb{R}^{r \times q}}{\arg\min} \left\{ \sum_{j=1}^q \|\boldsymbol{Y}_{\cdot,j} - \boldsymbol{X}\boldsymbol{U}\boldsymbol{c}_j\|_2^2 + \gamma \|\boldsymbol{U}\|_{1,2} \right\} \text{ subject to } \boldsymbol{C}\boldsymbol{C}^\top = I_r, \tag{39}$$

for which they develop an efficient optimization algorithm. Here, both $r \in \{1, 2, 3, \ldots\}$ and $\gamma > 0$ are user-specified tuning parameters.

Now, let us consider applying this approach to the regression problem with $J$ distinct populations. Assuming the $j$th population has $n_{(j)}$ subjects, each with univariate and continuous response $y_{(j)i}$ for $i \in [n_{(j)}]$ and $p$-dimensional predictor $\boldsymbol{x}_{(j)i}$, the analog of their estimator would be

$$\underset{\boldsymbol{U} \in \mathbb{R}^{p \times r}, \boldsymbol{C} \in \mathbb{R}^{r \times q}}{\arg\min} \left\{ \sum_{j=1}^J \|\boldsymbol{y}_{(j)} - \boldsymbol{X}_{(j)}\boldsymbol{U}\boldsymbol{c}_j\|_2^2 + \lambda \|\boldsymbol{U}\|_{1,2} \right\}, \quad \boldsymbol{C}\boldsymbol{C}^\top = I_r, \tag{40}$$

where $\boldsymbol{y}_{(j)} = (y_{(j)1}, \ldots, y_{(j)n_{(j)}})^\top$ and $\boldsymbol{X}_{(j)} = (\boldsymbol{x}_{(j)1}, \ldots, \boldsymbol{x}_{(j)n_{(j)}})^\top \in \mathbb{R}^{n_{(j)} \times p}$ for $j \in [J]$. The estimator in (40) can be characterized as (39) only if $\boldsymbol{X}_{(j)} = \boldsymbol{X}_{(j')} = \boldsymbol{X}$ for all $j \neq j'$. The

optimization problem in (40) is especially challenging relative to (39) because each component of the residual sum-of-squares depends on a distinct design matrix $\boldsymbol{X}_{(j)}$, so the computational approach developed in Chen and Huang (2012) cannot be applied directly. The same basic issue arises in applying the method of She (2017) in this context. Moreover, neither of their theoretical results apply to (40) in general.

Extending these approaches to handle $J$ populations with censored survival outcomes leads to further complication and would also require a new algorithm and theoretical framework.

# I Details on low rank decomposition of $\boldsymbol{B}_*$

In Section 6.2 of the main manuscript, we claim that when $J = 18$, $s = 20$, and $r = 6$, there are 192 parameters to be estimated. This follows from the fact that for a rank $r$ matrix $\boldsymbol{A} \in \mathbb{R}^{a\times b}$, the decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{a\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{b\times r}$, is nonunique. To make such a decomposition identifiable, it suffices to define

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_0 \\ I_r \end{pmatrix}$$

where $\boldsymbol{V}_0 \in \mathbb{R}^{(b-r)\times r}$ is unconstrained. This way, one need only estimate $\boldsymbol{U}$ and $\boldsymbol{V}_0$ so that there are $ar + (b-r)r = ar + br - r^2$ parameters needed to define $\boldsymbol{A}$. A more rigorous proof of this result—that an $a \times b$ matrix with rank $r$ is defined by $ar + br - r^2$ parameters—relies on the fact that such a matrix has $r$ nonzero singular values, and $r$ distinct pairs of singular vectors (which are subject to orthogonality and unit length constraints in $\mathbb{R}^a$ and $\mathbb{R}^b$). For example, see Section 2.3 of Velu and Reinsel (2013).

In our context, since $J = 18$, $r = 6$, and $s = 20$, this means we need only estimate the parameters of a rank $r$ matrix $\boldsymbol{B}_{\mathcal{S},\cdot} \in \mathbb{R}^{s\times J}$, the submatrix of $\boldsymbol{B}$ whose rows are nonzero. By the logic above, this means we need only estimate $sr + Jr - r^2 = 20 \cdot 6 + 18 \cdot 6 - 6^2 = 192$ parameters.

As a simple example, take $\boldsymbol{B}$, the $6 \times 5$ matrix with rank 3 given by

$$\boldsymbol{B} = \begin{pmatrix} -1.19 & -0.33 & -1.45 & -0.50 & -1.00 \\ -0.26 & -0.36 & -2.88 & 0.26 & -1.93 \\ -2.37 & 0.33 & 0.73 & -0.86 & -0.47 \\ 1.94 & 0.74 & 1.42 & 1.30 & 0.55 \\ -2.73 & 1.13 & 0.20 & 0.02 & -2.21 \\ -1.90 & 0.15 & 0.51 & -0.80 & -0.27 \end{pmatrix}.$$

Based on the discussion above, we can express this matrix in terms of $\boldsymbol{U} \in \mathbb{R}^{5\times 3}$ and $\boldsymbol{V}_0 \in \mathbb{R}^{2\times 3}$ as $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{V}^\top = \boldsymbol{U}[\boldsymbol{V}_0^\top, I_r] = [\boldsymbol{U}\boldsymbol{V}_0^\top, \boldsymbol{U}]$, from which is it easy to see $\boldsymbol{U} = \boldsymbol{B}_{*[,3:5]}$, and $\boldsymbol{V}_0 = \boldsymbol{U}^+\boldsymbol{B}_{*[,1:2]}$ so that

$$\boldsymbol{U} = \begin{pmatrix} -1.45 & -0.50 & -1.00 \\ -2.88 & 0.26 & -1.93 \\ 0.73 & -0.86 & -0.47 \\ 1.42 & 1.30 & 0.55 \\ 0.20 & 0.02 & -2.21 \\ 0.51 & -0.80 & -0.27 \end{pmatrix}, \quad \boldsymbol{V}_0 = \begin{pmatrix} -0.57 & 0.46 \\ 1.61 & 0.26 \\ 1.20 & -0.47 \end{pmatrix}.$$

# J Additional figures

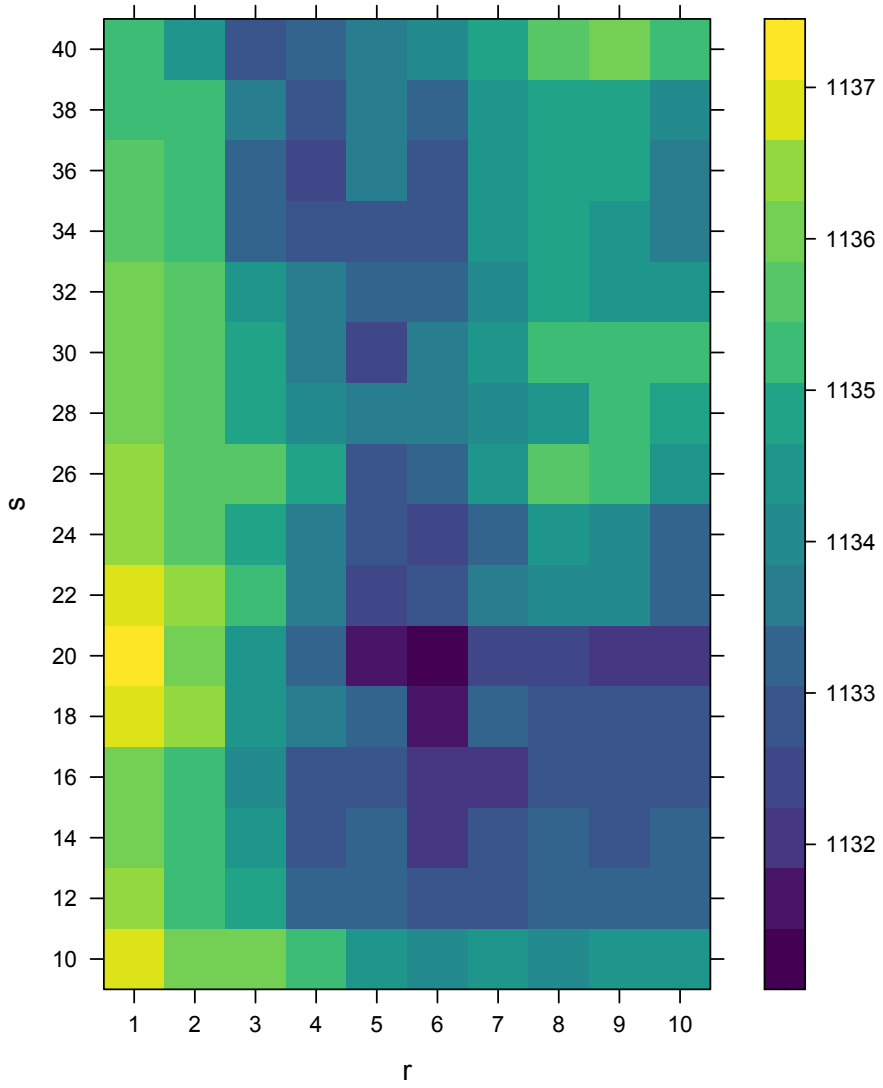


Figure 11: Five-fold cross-validation linear predictor scores from the real data analyzed in Section 6.

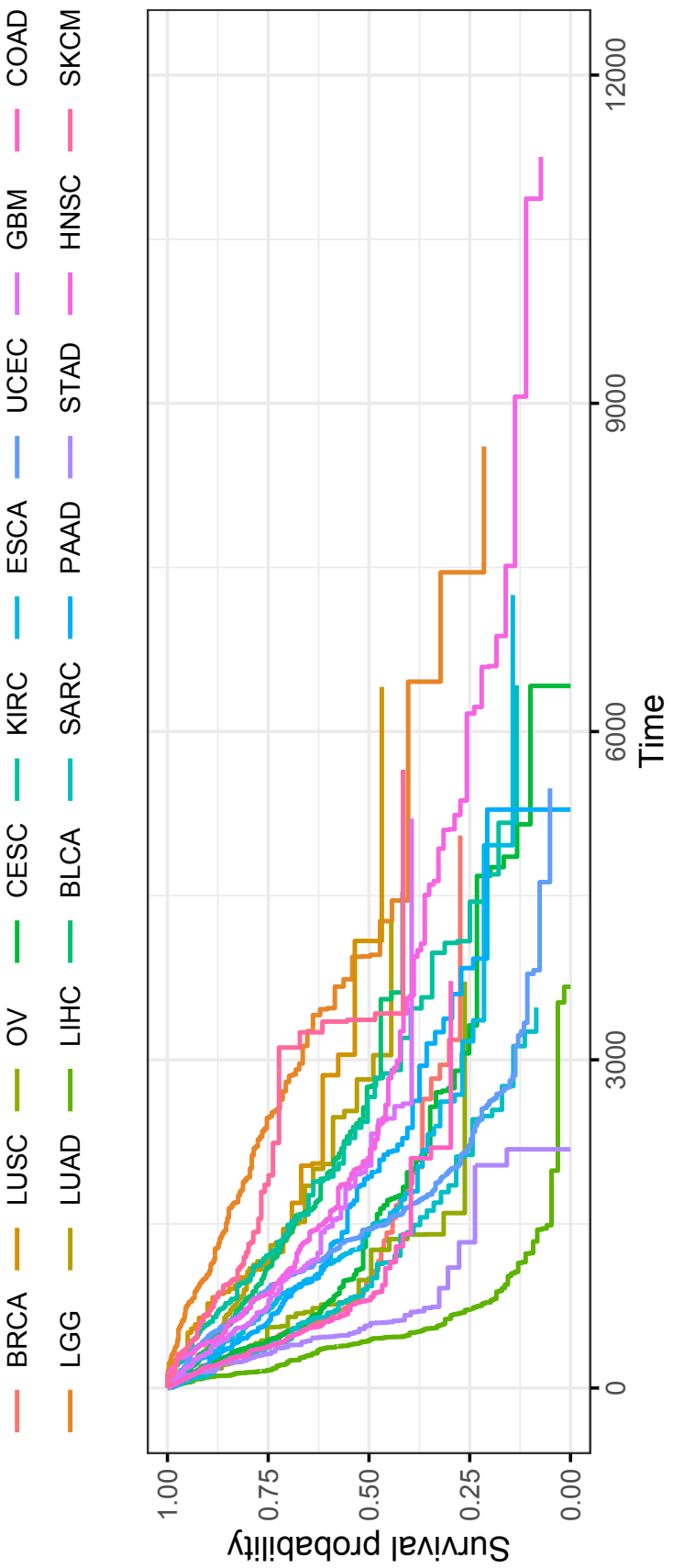


Figure 12: Kaplan-Meier survival curves for the 18 cancer datasets analyzed in Section 6.

# References


Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics*, 10(4):1100–1120.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545.

Dai, B. and Breheny, P. (2019). Cross validation approaches for penalized cox regression. https://arxiv.org/abs/1905.10432.

Davis, D. and Yin, W. (2017). A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis*, 25(4):829–858.

Hjort, N. L. and Pollard, D. (2011). Asymptotics for minimisers of convex processes. https://arxiv.org/abs/1107.3806.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239.

Pedregosa, F. and Gidel, G. (2018). Adaptive three operator splitting. In *International Conference on Machine Learning*, pages 4085–4094. PMLR.

Ryu, E. K. and Yin, W. (2019). Proximal-proximal gradient method. *Journal of Computational Mathematics*, 37(6):778–812.

She, Y. (2017). Selective factor extraction in high dimensions. *Biometrika*, 104(1):97–110.

Van der Vaart, A. (2002). Semiparametric statistics. In *Lectures on probability theory and statistics*, volume 1781 of *Lecture Notes in Math.*, pages 331–457. Springer.

Velu, R. and Reinsel, G. C. (2013). *Multivariate reduced-rank regression: Theory and applications*, volume 136. Springer Science & Business Media.

Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):463–483.