View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Caltech Authors - Main

# Factor analysis for survival time prediction with informative censoring and diverse covariates

## Supporting Information

Shannon R. McCurdy[2,1] Annette Molinaro[3] Lior Pachter[4]

[2] California Institute for Quantitative Biosciences,
University of California, Berkeley,
Berkeley, CA.

[3] Departments of Neurological Surgery, Epidemiology, and Biostatistics,
University of California, San Francisco,
San Francisco, CA.

[4] Division of Biology and Biological Engineering and
Department of Computing and Mathematical Sciences,
California Institute of Technology,
Pasadena, CA.

## Contents



[1] Corresponding author. E-mail: smccurdy@berkeley.edu

# SI-1 The exponential proportional hazards model with non-informative censoring (EPH-C) and $L_1$ penalty (EPH-C-$L_1$)

Let $t$, a non-negative random variable, be the survival time of an individual. We denote the non-random $(p \times 1)$-dimensional vector of covariates $\mathbf{x}$ for that individual. If a subset $s' = s + 1$ of the $p' = p + 1$ covariates have a fixed sum (e.g. categorical covariates), we only include $s$ elements of the subset in the $p$ covariates. Under the EPH model, there are two parameters of interest, the time-independent baseline hazard $\lambda$ and the $(p \times 1)$-dimensional vector of regression coefficients $\boldsymbol{\beta}$. For notational efficiency, we set,

$$\tilde{\mathbf{x}} = \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix} \qquad \mathbf{w} = \begin{pmatrix} \ln \lambda \\ \boldsymbol{\beta} \end{pmatrix} \tag{SI-1}$$

Under this model, the survival time distribution, given the covariates $\tilde{\mathbf{x}}$, is the exponential distribution with the following parameterization of the standard exponential rate parameter $\rho$,

$$t|\mathbf{x} \sim \text{Exp}(\rho = \exp\left(\mathbf{w}^T \tilde{\mathbf{x}}\right)) \tag{SI-2}$$

We take the following model for non-informative censoring (EPH-C). There are two non-negative random variables, $t$ and $c$, that represent the survival time and the censoring time

of the individual. We take the probability densities for $t$ and $c$ to be given by Eqn. (SI-2), with parameters $\mathbf{w}_T$ and $\mathbf{w}_C$, respectively, and $t \perp c|\mathbf{x}$. In the censored setting, one is not able to observe both $t$ and $c$ for an individual. Instead, the observed random variables for each individual are $\tilde{t} = \min(t, c)$ and $\delta = \mathbf{I}(\tilde{t} = t)$, where $\mathbf{I}$ is the indicator function. The probability of the observed data for an individual is simply,

$$p_{\mathbf{w}_T,\mathbf{w}_C}(\tilde{t}, \delta|\mathbf{x}) = \left(p_{\mathbf{w}_T}(\tilde{t}|\mathbf{x})P_{\mathbf{w}_C}(c > \tilde{t}|\mathbf{x})\right)^{\delta} \left(P_{\mathbf{w}_T}(t > \tilde{t}|\mathbf{x})p_{\mathbf{w}_C}(\tilde{t}|\mathbf{x})\right)^{1-\delta}. \tag{SI-3}$$

$P_{\mathbf{w}_{T,C}}((t, c) > \tilde{t}|\mathbf{x})$ is the survival function for $(t, c)$, respectively. As a shorthand, we use the subscript $\{T, C\}$ for generic expressions for the respective $(t, c)$ distributions.

We assume that the individual observations of the data are independent, so the likelihood of the data of $N$ individuals is simply the product of the individual probabilities. The likelihood of the data can be rearranged into a product of a partial likelihoods for $\mathbf{w}_T$ and $\mathbf{w}_C$,

$$\begin{aligned} \mathcal{L}(\mathbf{w}_T, \mathbf{w}_C) &= \mathcal{L}(\mathbf{w}_T)\mathcal{L}(\mathbf{w}_C) \\ \mathcal{L}(\mathbf{w}_{T,C}) &= \prod_{n=1}^{N} p_{\mathbf{w}_{T,C}}(\tilde{t}_n|\mathbf{x}_n)^{\delta_n^{T,C}} P_{\mathbf{w}_{T,C}}((t, c) > \tilde{t}_n|\mathbf{x}_n)^{1-\delta_n^{T,C}} \end{aligned} \tag{SI-4}$$

where we have introduced $\delta^T = \delta$ and $\delta^C = 1 - \delta$.

For PCA-EPH-C, the model is the same as EPH-C, except the non-random $(p \times 1)$-dimensional vector of covariates $\mathbf{x}$ is replaced with $(k \times 1)$-dimensional PCA projection of the covariates $\mathbf{x}$. Let the covariates $\mathbf{x}$ for $N$ individuals be arranged into a $(p \times N)$ mean-subtracted matrix with singular value decomposition $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$. Let $\mathbf{\Lambda}_k$ refer to the square matrix with the top $k$ singular values along the diagonal, and let $\mathbf{V}_k$ refer to the truncated $k \times N$ right singular vectors. Then the $(k \times N)$-dimensional covariates for PCA-EPH-C for individuals $1, \ldots, N$ are $\mathbf{\Lambda}_k\mathbf{V}_k$.

For EPH-C-$L_1$, the model is the same as EPH-C, except an $L_1$-penalty is desired, or, for the setting of $p > N$, required. For EPH-C-$L_1$, the partial log-likelihood of EPH-C gains a penalty factor,

$$\ln\left(\mathcal{L}_{L_1}(\mathbf{w}_{T,C})\right) = \ln\left(\mathcal{L}(\mathbf{w}_{T,C})\right) - \gamma_{T,C}\left|\mathbf{w}_{T,C}\right|, \tag{SI-5}$$

where $\gamma_{T,C}$ sets the degree of sparsity induced by the penalty. Determining the appropriate $\gamma_{T,C}$ is a model selection problem.

### SI-1.1 Inference

The parameters in the EPH model can be estimated through an iterative maximum likelihood procedure. Following the insight by [1], the maximum likelihood estimates,

$$\hat{\mathbf{w}}_{T,C} = \max_{\mathbf{w}_{T,C}} \mathcal{L}(\mathbf{w}_{T,C})$$

can be re-framed as an iterative least-squares minimization problem through a first-order Taylor expansion of $\mathcal{L}(\mathbf{w}_{T,C})$. For notational convenience, arrange the data for each individual $\tilde{\mathbf{x}}_n$ into a $(p+1) \times N$-dimensional matrix $\tilde{\mathbf{X}}$, and $\delta_n, \tilde{t}_n$ into $1 \times N$-dimensional vectors $\boldsymbol{\delta}, \tilde{\mathbf{t}}$. Another convenient definition is $\boldsymbol{\eta}_{T,C} = \mathbf{w}_{T,C}^T \tilde{\mathbf{X}}$. The minimization problem for $\hat{\mathbf{w}}_{T,C}^{(s+1)}$ at step $s+1$ is,

$$\hat{\mathbf{w}}_{T,C}^{(s+1)} = \min_{\mathbf{w}_{T,C}^{(s+1)}} \left|\left| \left( \boldsymbol{\eta}_{T,C}^{(s)} + \boldsymbol{\delta}^{T,C} \circ \tilde{\mathbf{t}}^{-1} \circ \exp\left(-\boldsymbol{\eta}_{T,C}^{(s)}\right) - \mathbb{1} - \left(\mathbf{w}_{T,C}^{(s+1)}\right)^T \tilde{\mathbf{X}} \right) \text{diag}\left( \sqrt{\tilde{\mathbf{t}}} \circ \exp(\tfrac{1}{2}\boldsymbol{\eta}_{T,C}^{(s)}) \right) \right|\right|^2$$

where $\circ$ is the element-wise Hadamard product, $(\exp(\mathbf{x}), \mathbf{x}^{-1})$ are applied element-wise, and $\mathbb{1}$ is a $1 \times N$ vector of ones. We initialize as follows,

$$\mathbf{w}_{T,C}^{(0)} = \begin{pmatrix} \ln\left( \frac{\sum_{n=1}^N \delta_n^{T,C}}{\sum_{n=1}^N \tilde{t}_n} \right) \\ \mathbb{0} \end{pmatrix},$$

where $\mathbb{0}$ is a $p \times 1$ vector of zeros.

If an $L_1$ penalized solution for $\hat{\mathbf{w}}_{T,C}$ is desired, use standard algorithms such as least-angle-regression-lasso (LARS-lasso) [2] at each iteration [1].

We find $\hat{\mathbf{w}}_{T,C}$ converges quickly; we use 5 iterations.

### SI-1.2 Prediction

Once we have the maximum likelihood estimate for the regression coefficients, $\hat{\mathbf{w}}_T$, we can calculate the expectation of the time-to-event given the covariates $\mathbf{x}$ for an individual. Under the EPH model (Eqn. SI-2), the prediction $\hat{t}$ for the individual's time-to-event is,

$$\hat{t} = \mathbb{E}_{\hat{\mathbf{w}}_T}[t|\mathbf{x}] = \exp(-\hat{\mathbf{w}}_T^T \tilde{\mathbf{x}}).$$

For PCA-EPH-C prediction, the $(k \times 1)$-dimensional PCA projection of the covariates for the an unseen sample $\mathbf{x}$ is, $\mathbf{U}_k^T(\mathbf{x} - \boldsymbol{\mu})$, where the $\mathbf{U}_k^T$ and $\boldsymbol{\mu}$ are learned from the training samples.

## SI-2 Factor analysis (FA)

### SI-2.1 Inference

There is no closed-form solution for the maximum likelihood estimates of the parameters of FA, but the estimates can be obtained through the Expectation-Maximization (EM) algorithm [3] applied to the complete log-likelihood [4]. For notational convenience, arrange the data for each individual $\mathbf{x}_n$ into a $d_x \times N$-dimensional matrix $\mathbf{X}$, $\mathbf{z}_n$ into a $d_z \times N$-dimensional matrix $\mathbf{Z}$, and $\boldsymbol{\mu}$ is a $d_x \times N$ matrix with each column identical. We follow [4] and find $\hat{\boldsymbol{\mu}}$ from the maximum likelihood of the marginal likelihood of the observed covariates,

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n. \tag{SI-6}$$

From the complete log-likelihood, then, the expectation (E) -step, given parameters $\boldsymbol{\mu}, \mathbf{W}, \boldsymbol{\Psi}$ is,

$$\mathbb{E}[\mathbf{Z}|\mathbf{X}] = (\mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W} + \mathbb{1})^{-1}\mathbf{W}^T\boldsymbol{\Psi}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \tag{SI-7}$$

$$\mathbb{E}[\mathbf{Z}\mathbf{Z}^T|\mathbf{X}] = N(\mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W} + \mathbb{1})^{-1} + \mathbb{E}[\mathbf{Z}|\mathbf{X}]\mathbb{E}[\mathbf{Z}|\mathbf{X}]^T. \tag{SI-8}$$

The maximization (M) -step of the parameters is,

$$\begin{aligned}\hat{\mathbf{W}}|\hat{\boldsymbol{\mu}} &= (\mathbf{X} - \hat{\boldsymbol{\mu}})\,\mathbb{E}[\mathbf{Z}|\mathbf{X}]^T\mathbb{E}[\mathbf{Z}\mathbf{Z}^T|\mathbf{X}]^{-1} \\ \hat{\boldsymbol{\Psi}}|\hat{\boldsymbol{\mu}}, \hat{\mathbf{W}} &= \frac{1}{N}\mathrm{diag}\left((\mathbf{X} - \hat{\boldsymbol{\mu}})(\mathbf{X} - \hat{\boldsymbol{\mu}})^T - \hat{\mathbf{W}}\mathbb{E}[\mathbf{Z}\mathbf{Z}^T|\mathbf{X}]\hat{\mathbf{W}}^T\right).\end{aligned} \tag{SI-9}$$

#### SI-2.1.1 Additional conditional distributions

The conditional binomial distribution, Binomial$(b, \mathbf{f} = \sigma(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}))$, is,

$$p_{b,\mathbf{W},\boldsymbol{\mu}}(\mathbf{x}|\mathbf{z}) = \left(\prod_{i=1}^{d_x}\binom{b}{\mathbf{x}_i}\sigma^b(-(\mathbf{W}_i\mathbf{z} + \boldsymbol{\mu}_i))\right)\exp\left(\mathbf{x}^T\left(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}\right)\right) \tag{SI-10}$$

where the elements of $\mathbf{x}$ are $\mathbf{x}_i \in \{0, 1, \ldots, b\}$. This conditional distribution has a drawback in that the marginal distribution of the observed covariates, $p(\mathbf{x})$, is intractable. Using the insight of [5] and [6], we introduce a variational approximation to the logistic function $\sigma$ in (Eqn. SI-10). This approximation is,

$$\sigma(x) \geq \sigma(\xi)\exp\left(\frac{1}{2}(x - \xi) - \lambda(\xi)(x^2 - \xi^2)\right) \qquad \lambda(\xi) = \frac{1}{2\xi}\left(\sigma(\xi) - \frac{1}{2}\right), \tag{SI-11}$$

where $\xi$ is the variational parameter. Note that the approximation is exact when $\xi = x$. Under this approximation, $p_{b,\mathbf{W},\boldsymbol{\mu}}(\mathbf{x}|\mathbf{z})$ is approximated by $\tilde{p}_{b,\mathbf{W},\boldsymbol{\mu},\boldsymbol{\xi}}(\mathbf{x}|\mathbf{z})$, where there is a $d_x \times 1$ vector of variational parameters $\boldsymbol{\xi}$ (for $N$ individuals there are $d_x \times N$),

$$\begin{aligned}\tilde{p}_{b,\mathbf{W},\boldsymbol{\mu},\boldsymbol{\xi}}(\mathbf{x}|\mathbf{z}) &= \left(\prod_{i=1}^{d_x}\binom{b}{x_i}\sigma^b(\boldsymbol{\xi}_i)\exp\left(-\frac{b}{2}(\mathbf{W}_i\mathbf{z}+\boldsymbol{\mu}_i+\boldsymbol{\xi}_i)-b\lambda(\boldsymbol{\xi}_i)((\mathbf{W}_i\mathbf{z}+\boldsymbol{\mu}_i)^2-\boldsymbol{\xi}_i^2)\right)\right) \\ &\quad \exp\left(\mathbf{x}^T(\mathbf{W}\mathbf{z}+\boldsymbol{\mu})\right). \qquad \text{(SI-12)}\end{aligned}$$

This approximation is a lower bound: $\tilde{p}_{b,\mathbf{W},\boldsymbol{\mu},\boldsymbol{\xi}}(\mathbf{x}|\mathbf{z}) \leq p_{b,\mathbf{W},\boldsymbol{\mu}}(\mathbf{x}|\mathbf{z})$ and $\tilde{p}_{b,\mathbf{W},\boldsymbol{\mu},\boldsymbol{\xi}}(\mathbf{x}) \leq p_{b,\mathbf{W},\boldsymbol{\mu}}(\mathbf{x})$. An additional benefit is that the variational approximation is quadratic in $\mathbf{z}$. As a consequence, the EM updates for the approximate complete log-likelihood are analytic, and the conditional probability of $\tilde{p}(\mathbf{z}|\mathbf{x})$ becomes Gaussian.

With this approximation, for individuals $n \in \{1,\ldots,N\}$, the E-steps are,

$$\mathbb{E}[\mathbf{z}_{jn}|\mathbf{x}_n] = \sum_{k=1}^{d_z}(\mathbf{C}_n)_{jk}\left(\sum_{i=1}^{d_x}\mathbf{W}_{ik}\left(\mathbf{x}_{in}-\frac{b}{2}-2b\lambda(\boldsymbol{\xi}_{in})\boldsymbol{\mu}_{in}\right)\right) \qquad \text{(SI-13)}$$

$$\begin{aligned}\mathbb{E}[\mathbf{z}_{jn}\mathbf{z}_{kn}|\mathbf{x}_n] &= (\mathbf{C}_n)_{jk}+\mathbb{E}[\mathbf{z}_{jn}|\mathbf{x}_n]\mathbb{E}[\mathbf{z}_{kn}|\mathbf{x}_n] \\ &\quad \text{where } ((\mathbf{C}_n)^{-1})_{jk} = \left(\delta_{jk}+2b\sum_{i=1}^{d_x}\lambda(\boldsymbol{\xi}_{in})\mathbf{W}_{ij}\mathbf{W}_{ik}\right) \qquad \text{(SI-14)}\end{aligned}$$

In these expressions, repeated indices do not imply summations. $\boldsymbol{\mu}_{in}$ is the $(in)^{th}$ element of the matrix of $N$ columns of the $d_x \times 1$ vector $\boldsymbol{\mu}$. $\delta_{jk}$ is the Kronecker delta. Each $\mathbf{C}_n$ is a $d_z \times d_z$ matrix.

The M-steps are,

$$\begin{aligned}\hat{\boldsymbol{\xi}}_{in}^2|\mathbf{W}_{ij},\boldsymbol{\mu}_{in} &= \sum_{k=1}^{d_z}\sum_{j=1}^{d_z}\mathbf{W}_{ij}\mathbf{W}_{ik}\mathbb{E}[\mathbf{z}_{jn}\mathbf{z}_{kn}|\mathbf{x}_n]+2\sum_{j=1}^{d_z}\mathbf{W}_{ij}\mathbb{E}[\mathbf{z}_{jn}|\mathbf{x}_n]\boldsymbol{\mu}_{in}+\boldsymbol{\mu}_{in}^2 \\ \hat{\mathbf{W}}_{ik}|\hat{\boldsymbol{\xi}}_{in},\boldsymbol{\mu}_{in} &= \min_{\mathbf{W}_{ik}}\frac{1}{2}\left(\sum_{j=1}^{d_z}\sum_{n=1}^{N}\left(\mathbf{x}_{in}-\frac{b}{2}-2b\lambda(\hat{\boldsymbol{\xi}}_{in})\boldsymbol{\mu}_{in}\right)\mathbb{E}[\mathbf{z}_{jn}|\mathbf{x}_n]((\mathbf{L}_i)^{-1T})_{jk}\right. \\ &\quad \left.-\sum_{j=1}^{d_z}\mathbf{W}_{ij}(\mathbf{L}_i)_{jk}\right)^2 \\ &\quad \text{where } \sum_{l=1}^{d_z}(\mathbf{L}_i)_{jl}(\mathbf{L}_i)_{lk}^T = \sum_{n=1}^{N}2b\lambda(\hat{\boldsymbol{\xi}}_{in})\mathbb{E}[\mathbf{z}_{jn}\mathbf{z}_{kn}|\mathbf{x}_n] \\ \hat{\boldsymbol{\mu}}_i|\hat{\boldsymbol{\xi}}_{in},\hat{\mathbf{W}}_{ij} &= \frac{1}{\sum_{n=1}^{N}2b\lambda(\hat{\boldsymbol{\xi}}_{in})}\sum_{n=1}^{N}\left(\left(\mathbf{x}_{in}-\frac{b}{2}\right)-2b\lambda(\hat{\boldsymbol{\xi}}_{in})\sum_{j=1}^{d_z}\hat{\mathbf{W}}_{ij}\mathbb{E}[\mathbf{z}_{jn}|\mathbf{x}_n]\right) \qquad \text{(SI-15)}\end{aligned}$$

where each $(\mathbf{L}_i)$ is the $i^{th}$ (in $d_x$) $(d_z \times d_z)$-Cholesky decomposition. There are $d_x$ total $d_z$-dimensional minimization problems for $\hat{\mathbf{W}}_i$. In these expressions, repeated indices do not

imply summations. We have revealed the indices and summations for the sake of clarity. Note that this is an expectation-conditional-maximization algorithm (ECM) due to the structure of successive maximizations [7]. We have suppressed a EM-iteration index in favor of $\hat{\cdot}$ referring to the current iteration's maximum likelihood estimate of the parameter and the absence of a hat $\cdot$ on a parameter referring to the prior iteration's maximum likelihood estimate of the parameter.

The conditionally multinomial distribution Multinomial$(b, \mathbf{f} = \text{softmax}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}))$ is,

$$p_{b,\mathbf{W},\boldsymbol{\mu}}(\mathbf{x}|\mathbf{z}) = \frac{b!}{\prod_{i=1}^{d_x} \mathbf{x}_i!} \exp\left(\mathbf{x}^T(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}) - b \ln\left(\sum_{i=1}^{d_x} \exp\left(\mathbf{W}_i\mathbf{z} + \boldsymbol{\mu}_i\right)\right)\right) \tag{SI-16}$$

where $b = \sum_{i=1}^{d_x} \mathbf{x}_i$ and $\sum_{i=1}^{d_x} \mathbf{f}_i = 1$. To have the appropriate number of parameters for the multinomial distribution, we set the elements of the last row of $\mathbf{W}$ and $\boldsymbol{\mu}$ equal to zero.

[8] introduce the following two-step bound on the function $\ln\left(\sum_{i=1}^{d_x} \exp\left(\boldsymbol{\eta}_i\right)\right)$. This bound builds on the work of [5] and [6] and allows for a closed-form approximate EM algorithm for the multinomial factor analysis setting. The bound in [8] is,

$$\ln\left(\sum_{i=1}^{d_x} \exp(\boldsymbol{\eta}_i)\right) \leq \alpha + \sum_{i=1}^{d_x} \ln(1 + \exp(\boldsymbol{\eta}_i - \alpha)) = \alpha - \sum_{i=1}^{d_x} \ln \sigma(-\boldsymbol{\eta}_i + \alpha) \tag{SI-17}$$

Next, the variational approximation (Eqn. SI-11) is applied separately to each $\sigma(-\boldsymbol{\eta}_i + \alpha)$. This is the same as $i$ independent binomial approximations, with one more parameter, $\alpha$, for the overall constraint. The approximate conditional multivariate distribution is,

$$\begin{aligned} \tilde{p}_{b,\mathbf{W},\boldsymbol{\mu},\boldsymbol{\xi}}(\mathbf{x}|\mathbf{z}) &= \frac{b!}{\prod_{i=1}^{d_x} \mathbf{x}_i!} \exp\left(-b\alpha\right) \sigma^b(\boldsymbol{\xi}_i) \exp\left(\mathbf{x}^T(\mathbf{W}\mathbf{z} + \boldsymbol{\mu})\right) \\ &\quad \prod_{i=1}^{d_x} \exp\left(-\frac{b}{2}(\mathbf{W}_i\mathbf{z} + \boldsymbol{\mu}_i - \alpha + \boldsymbol{\xi}_i) - b\lambda(\boldsymbol{\xi}_i)((\mathbf{W}_i\mathbf{z} + \boldsymbol{\mu}_i - \alpha)^2 - \boldsymbol{\xi}_i^2)\right) \end{aligned} \tag{SI-18}$$

For $N$ individuals, $\boldsymbol{\alpha}$ is a $1 \times N$ vector, and, as in the binomial case, there are $d_x \times N$ variational parameters $\boldsymbol{\xi}$.

Like the binomial case, this approximation provides a lower bound: $\tilde{p}_{b,\mathbf{W},\boldsymbol{\mu},\boldsymbol{\xi}}(\mathbf{x}|\mathbf{z}) \leq p_{b,\mathbf{W},\boldsymbol{\mu}}(\mathbf{x}|\mathbf{z})$ and $\tilde{p}_{b,\mathbf{W},\boldsymbol{\mu},\boldsymbol{\xi}}(\mathbf{x}) \leq p_{b,\mathbf{W},\boldsymbol{\mu}}(\mathbf{x})$. The variational approximation is again quadratic in $\mathbf{z}$, and the EM updates for the approximate complete log-likelihood are analytic, and the conditional probability of $\tilde{p}(\mathbf{z}|\mathbf{x})$ becomes Gaussian.

With this approximation, for individuals $n \in \{1, \ldots, N\}$, the E-steps are,

$$\begin{aligned} \mathbb{E}[\mathbf{z}_{jn}|\mathbf{x}_n] &= \sum_{k=1}^{d_z} (\mathbf{C}_n)_{jk} \left(\sum_{i=1}^{d_x} \mathbf{W}_{ik} \left(\mathbf{x}_{in} - \frac{b}{2} - 2b\lambda(\boldsymbol{\xi}_{in})(\boldsymbol{\mu}_{in} - \boldsymbol{\alpha}_n)\right)\right) \\ \mathbb{E}[\mathbf{z}_{jn}\mathbf{z}_{kn}|\mathbf{x}_n] &= (\mathbf{C}_n)_{jk} + \mathbb{E}[\mathbf{z}_{jn}|\mathbf{x}_n]\mathbb{E}[\mathbf{z}_{kn}|\mathbf{x}_n] \end{aligned} \tag{SI-19}$$

$$\text{where } ((\mathbf{C}_n)^{-1})_{jk} = \left(\delta_{jk} + 2b\sum_{i=1}^{d_x}\lambda(\boldsymbol{\xi}_{in})\mathbf{W}_{ij}\mathbf{W}_{ik}\right) \tag{SI-20}$$

In these expressions, repeated indices do not imply summations. $\boldsymbol{\mu}_{in}$ is the $(in)^{th}$ element of the matrix of $N$ columns of the $d_x \times 1$ vector $\boldsymbol{\mu}$. $\delta_{jk}$ is the Kronecker delta. Each $\mathbf{C}_n$ is a $d_z \times d_z$ matrix.

The M-steps are,

$$\begin{aligned}
\hat{\boldsymbol{\xi}}_{in}^2|\mathbf{W}_{ij},\boldsymbol{\mu}_{in},\boldsymbol{\alpha}_n &= \sum_{k=1}^{d_z}\sum_{j=1}^{d_z}\mathbf{W}_{ij}\mathbf{W}_{ik}\mathbb{E}[\mathbf{z}_{jn}\mathbf{z}_{kn}|\mathbf{x}_n] + 2\sum_{j=1}^{d_z}\mathbf{W}_{ij}\mathbb{E}[\mathbf{z}_{jn}|\mathbf{x}_n]\boldsymbol{\mu}_{in} + \boldsymbol{\mu}_{in}^2 \\
&\quad -2\boldsymbol{\alpha}_n\sum_{j=1}^{d_z}\mathbf{W}_{ij}\mathbb{E}[\mathbf{z}_{jn}|\mathbf{x}_n] + \boldsymbol{\alpha}_n^2 - 2\boldsymbol{\alpha}_n\boldsymbol{\mu}_{in} \\
\hat{\boldsymbol{\alpha}}_n|\hat{\boldsymbol{\xi}}_{in},\mathbf{W}_{ij},\boldsymbol{\mu}_{in} &= \frac{1}{\sum_{i=1}^{d_x}\lambda(\hat{\boldsymbol{\xi}}_{in})}\left(\sum_{i=1}^{d_x}\lambda(\hat{\boldsymbol{\xi}}_{in})\left(\sum_{j=1}^{d_z}\mathbf{W}_{ij}\mathbb{E}[\mathbf{z}_{jn}|\mathbf{x}_n] + \boldsymbol{\mu}_{in}\right) - \frac{1-\frac{d_x}{2}}{2}\right) \\
\hat{\mathbf{W}}_{ik}|\hat{\boldsymbol{\xi}}_{in},\hat{\boldsymbol{\alpha}}_n,\boldsymbol{\mu}_{in} &= \min_{\mathbf{W}_{ik}}\frac{1}{2}\left(\sum_{j=1}^{d_z}\sum_{n=1}^{N}\left(\mathbf{x}_{in} - \frac{b}{2} - 2b\lambda(\hat{\boldsymbol{\xi}}_{in})\left(\boldsymbol{\mu}_{in} - \hat{\boldsymbol{\alpha}}_n\right)\right)\mathbb{E}[\mathbf{z}_{jn}|\mathbf{x}_n]((\mathbf{L}_i)^{-1T})_{jk}\right. \\
&\quad \left. -\sum_{j=1}^{d_z}\mathbf{W}_{ij}(\mathbf{L}_i)_{jk}\right)^2 \\
&\quad \text{where } \sum_{l=1}^{d_z}(\mathbf{L}_i)_{jl}(\mathbf{L}_i)_{lk}^T = \sum_{n=1}^{N}2b\lambda(\hat{\boldsymbol{\xi}}_{in})\mathbb{E}[\mathbf{z}_{jn}\mathbf{z}_{kn}|\mathbf{x}_n] \\
\hat{\boldsymbol{\mu}}_i|\hat{\boldsymbol{\xi}}_{in},\hat{\mathbf{W}}_{ij},\hat{\boldsymbol{\alpha}}_n &= \frac{1}{\sum_{n=1}^{N}2b\lambda(\hat{\boldsymbol{\xi}}_{in})}\sum_{n=1}^{N}\left(\left(\mathbf{x}_{in} - \frac{b}{2}\right) - 2b\lambda(\hat{\boldsymbol{\xi}}_{in})\left(\sum_{j=1}^{d_z}\hat{\mathbf{W}}_{ij}\mathbb{E}[\mathbf{z}_{jn}|\mathbf{x}_n] - \hat{\boldsymbol{\alpha}}_n\right)\right),
\end{aligned} \tag{SI-21}$$

where the conventions are the same as for the conditionally binomial case. We choose an ECM approach for $\mathbf{W}$ and $\boldsymbol{\mu}$ for the same reasons as for the conditionally binomial case. We also choose an ECM approach for the variational parameters to avoid large matrix inversions.

### SI-2.1.2 Diverse conditional distributions

For diverse conditional distributions such as (Eqn. 12), with the variational approximations in place for the conditionally binomial covariates and conditionally multinomial covariates, the conditional probability of $\tilde{p}(\mathbf{z}|\mathbf{x})$ remains Gaussian. The E-step for individuals $n \in \{1,\ldots,N\}$ are,

$$\begin{aligned}
\mathbb{E}[\mathbf{z}_{jn}|\mathbf{x}_n] &= \sum_{k=1}^{d_z}(\mathbf{C}_n)_{jk}\sum_{d=1}^{D}\left(\mathbf{I}(\text{type}(d) = \text{normal})\sum_{i=1}^{d_x^{(d)}}\mathbf{W}_{ik}^{(d)}(\boldsymbol{\Psi}^{(d)})_{ii}^{-1}(\mathbf{x}_{in}^{(d)} - \boldsymbol{\mu}_{in}^{(d)})\right. \\
&\quad \left. +\mathbf{I}(\text{type}(d) = \text{binomial})\sum_{i=1}^{d_x^{(d)}}\mathbf{W}_{ik}^{(d)}\left(\mathbf{x}_{in}^{(d)} - \frac{b^{(d)}}{2} - 2b^{(d)}\lambda(\boldsymbol{\xi}_{in}^{(d)})\boldsymbol{\mu}_{in}^{(d)}\right)\right.
\end{aligned}$$

$$
\begin{aligned}
&+\mathbf{I}(\text{type}(d) = \text{multinomial}) \sum_{i=1}^{d_x^{(d)}} \mathbf{W}_{ik}^{(d)} \left( \mathbf{x}_{in}^{(d)} - \frac{b^{(d)}}{2} - 2b^{(d)} \lambda(\boldsymbol{\xi}_{in}^{(d)})(\boldsymbol{\mu}_{in}^{(d)} - \boldsymbol{\alpha}_n^{(d)}) \right) \Bigg) \\
\mathbb{E}[\mathbf{z}_{jn}\mathbf{z}_{kn}|\mathbf{x}_n] &= (\mathbf{C}_n)_{jk} + \mathbb{E}[\mathbf{z}_{jn}|\mathbf{x}_n]\mathbb{E}[\mathbf{z}_{kn}|\mathbf{x}_n] \\
&\text{where } ((\mathbf{C}_n)^{-1})_{jk} = \left( \delta_{jk} + \sum_{d=1}^{D} \left( \mathbf{I}(\text{type}(d) = \text{normal}) \sum_{i=1}^{d_x^{(d)}} \mathbf{W}_{ij}^{(d)} (\boldsymbol{\Psi}^{(d)})_{ii}^{-1} \mathbf{W}_{ik}^{(d)} \right. \right. \\
&\left. \left. + \mathbf{I}(\text{type}(d) \neq \text{normal}) \, 2b^{(d)} \sum_{i=1}^{d_x^{(d)}} \lambda(\boldsymbol{\xi}_{in}^{(d)}) \mathbf{W}_{ij}^{(d)} \mathbf{W}_{ik}^{(d)} \right) \right). \qquad \text{(SI-22)}
\end{aligned}
$$

The M-steps for each $(d)$ data type's parameters are simply the M-steps for the relevant data type given in (eqns. SI-9, SI-15, SI-21).

#### SI-2.1.3 Initialization

For each data type $(d)$, we take a warm start initialization of the EM algorithm. In the following, we will omit the data type index. If $d_x > d_z$, we initialize FA at the probabilistic PCA solution [9],

$$
\begin{aligned}
\boldsymbol{\mu}^{(0)} &= \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \\
\mathbf{W}^{(0)} &= \mathbf{U}_{d_z} (\boldsymbol{\Lambda}_{d_z}^2 - \sigma^2 \mathbb{1}) \\
\boldsymbol{\Psi}^{(0)} &= \sigma^2 \mathbb{1} \\
&\text{where } \mathbf{X} - \boldsymbol{\mu}^{(0)} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T \text{ is the singular value decomposition} \\
&\text{and } \sigma^2 = \frac{1}{N(d_x - d_z)} \sum_{i=d_z+1}^{d_x} \Lambda_{ii}^2 \qquad \text{(SI-23)}
\end{aligned}
$$

and $\mathbf{U}_{d_z}$ is the first $d_z$ columns of $\mathbf{U}$, and likewise for $\boldsymbol{\Lambda}$. If $d_x \leq d_z$, then we initialize $\mathbf{W}^{(0)}$ as a matrix of ones, and $\boldsymbol{\Psi}^{(0)} = (\Lambda_{ii}^2 \mathbf{I}(i = d_x)/N)\mathbb{1}$.

We initialize all variational parameters to one ($\boldsymbol{\alpha}_n = 1, \boldsymbol{\xi}_{in} = 1$), and enforce the appropriate constraints on $\mathbf{W}, \boldsymbol{\mu}$ for the conditionally multinomial distributions.

## SI-3 Joint factor analysis and exponential proportional hazards model with informative censoring (FA-EPH-C)

### SI-3.1 Inference

With the addition of time-to-event and censoring data, the marginal probability of the data $p_\theta(\tilde{t}, \delta, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)})$ is not analytic, even with the variational approximations. As a result, the conditional expectations of the latent variables given the data are also not analytic. We

calculate the necessary conditional expectations of functions $f(\mathbf{z})$ given $\tilde{t}, \delta, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}$ by sampling from the conditional distribution using the Metropolis-Hastings (MH) algorithm [10; 11]. The use of Monte-Carlo methods to approximate the E-step in the EM algorithm was first introduced by [12]. The proposal density we use is $\mathbf{z}'_n|\mathbf{z}_n^{(s)} \sim \mathcal{N}(\mathbf{z}_n^{(s)}, \kappa\,\mathbf{C}_n)$, where $\mathbf{z}'_n$ is the proposal sample for step $(s+1)$ and individual $n$, $\mathbf{C}_n$ is the covariance of individual $n$ under the FA-only model (Eqn. SI-22), and $\kappa$ is a scale parameter. We initialize at $\mathbf{z}_n^{(0)} = \mathbb{E}[\mathbf{z}_n|\mathbf{x}_n]$, the conditional expectation of individual $n$ under the FA-only model (Eqn. SI-22).

We discard the first $s = \{1, \ldots, 300\}$ burn-in samples and collect the $s = \{301, \ldots, 600\}$ samples for each individual $n$. We use these 300 samples for each individual $n$ to calculate the empirical conditional expectations $\mathbb{E}[f(\mathbf{z})|\tilde{t}, \delta, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}]$.

We monitor the acceptance rate $r_a$, effective sample size $n_{eff}$, and convergence parameter $\hat{R}(|\mathbb{E}[\mathbf{z}_n|\mathbf{x}_n]|^2)$ for two parallel Metropolis-Hastings sampling chains for the first E-step for the first-individual to tune the scale parameter $\kappa$ [13]. The effective sample size is defined in [13] (Eqn. 11.4), and the convergence parameter $\hat{R}$ of a statistic is defined in [13] between (eqns. 11.3-4). Starting with $\kappa = 6$ and continuing in descending order ( $\kappa \in \{6, 5.5, 5, 4.5, 4, 3.5, 3, 2.5, 2, 1.5, 1, 0.5, 0.25, 0.1\}$), we select the first scale parameter that has an acceptance rate between $0.134 \leq r_a \leq 0.334$, an effective sample size $n_{eff} \geq 10$, and a convergence parameter $\hat{R}(|\mathbb{E}[\mathbf{z}_n|\mathbf{x}_n]|^2) \leq 1.2$. [13] contains a comprehensive discussion of these statistics.

The M-step for each $\mathbf{x}^{(d)}$ data type's parameters is simply the M-step for the relevant data type given in (eqns. SI-9, SI-15, SI-21). For the time-to-event and censoring data type M-steps, we perform a Newton-Raphson step [14]. The subsequent EM algorithm is a generalized EM (GEM) algorithm [3]. The M-step is the same for $\mathbf{w}_{T,C}$ with $\delta^{T,C}$. We omit the $T, C$ index below. At step $(s+1)$, the M-step is,

$$\begin{aligned}
\mathbf{w}^{(s+1)} \quad = \quad & \mathbf{w}^{(s)} + a \left( \sum_{n=1}^{N} \tilde{t}_n \mathbb{E}_{\Theta^{(s)}} \left[ \tilde{\mathbf{z}}_n \tilde{\mathbf{z}}_n^T \exp\left( \left( \mathbf{w}^{(s)} \right)^T \tilde{\mathbf{z}}_n \right) \middle| \tilde{t}, \delta, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)} \right] \right)^{-1} \\
& \sum_{n=1}^{N} \left( \delta_n \mathbb{E}_{\Theta^{(s)}}[\tilde{\mathbf{z}}_n|\tilde{t}, \delta, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}] - \tilde{t}_n \mathbb{E}_{\Theta^{(s)}} \left[ \tilde{\mathbf{z}}_n \exp\left( \left( \mathbf{w}^{(s)} \right)^T \tilde{\mathbf{z}}_n \right) \middle| \tilde{t}, \delta, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)} \right] \right) \\
& \text{where} \qquad 0 < a \leq 1.
\end{aligned}$$

The parameter $a$ sets the scale of the Newton-Raphson step size, $\tilde{\mathbf{z}}$ is defined as in (Eqn. SI-1), and the expectations are taken with respect to the step $(s)$ parameters. We set $a = 1$.

We initialize at,

$$\mathbf{w}_{T,C}^{(0)} = \begin{pmatrix} \ln\left(\frac{\sum_{n=1}^{N}\delta_n^{T,C}}{\sum_{n=1}^{N}\tilde{t}_n}\right) \\ \mathbb{0} \end{pmatrix}. \tag{SI-24}$$

For the results in this note, we stop the approximate GEM algorithm after 10 iterations. We explored using a larger number of iterations but found no improvement in the simulated or cross-validated expected c-index (Sec. SI-6).

### SI-3.2 Prediction

Once we have estimates for the maximum likelihood parameters $\hat{\mathbf{w}}_T, \hat{\mathbf{w}}_C, \hat{\Theta}$, we would like to predict a survival time $\hat{t}$ given $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}$ and the maximum likelihood parameters $\hat{\mathbf{w}}_T, \hat{\mathbf{w}}_C, \hat{\Theta}$. With the variational approximations for binomial and multinomial distributions, the conditional expectation of the survival time $t$ is analytic. The result is,

$$\hat{t} = \mathbb{E}_{\hat{\mathbf{w}}_T,\hat{\mathbf{w}}_C,\hat{\Theta}}[t|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}] = \frac{1}{\hat{\lambda}_T} \exp\left(\frac{1}{2}\hat{\boldsymbol{\beta}}_T^T \hat{\mathbf{C}} \hat{\boldsymbol{\beta}}_T - \mathbb{E}_{\hat{\mathbf{w}}_T,\hat{\mathbf{w}}_C,\hat{\Theta}}[\mathbf{z}^T|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}]\hat{\boldsymbol{\beta}}_T\right) \tag{SI-25}$$

where $\hat{\mathbf{C}}$ and $\mathbb{E}_{\hat{\mathbf{w}}_T,\hat{\mathbf{w}}_C,\hat{\Theta}}[\mathbf{z}^T|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}]$ are defined in (Eqn. SI-22), and $\hat{\boldsymbol{\beta}}_T, \hat{\lambda}_T$ are components of $\hat{\mathbf{w}}_T$ in (Eqn. SI-1). For these predictions, we take the average variational parameter values $\hat{\alpha} = \frac{1}{N}\sum_{n=1}^{N}\hat{\boldsymbol{\alpha}}_n$ and $\hat{\boldsymbol{\xi}}_i = \frac{1}{N}\sum_{n=1}^{N}\hat{\boldsymbol{\xi}}_{in}$, since the estimates for the variational parameters $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\xi}}$ were dependent on the learning-set individual $n$.

## SI-4 Fast Approximation

We find a fast, decoupled approximation to the fully integrative model FA-EPH-C. The fast approximation is as follows: (1) inference for factor analysis and estimation of the conditional expectation of the latent variables, and then (2) inference for an exponential proportional hazards model with the conditional expectation of the latent variables as covariates. The decoupled inference algorithms are analytic. This provides a significant speed-up compared to the fully integrative model, which requires Metropolis-Hastings simulations for every individual at every E-step.

We tested the fast approximation on the cross validation stage of all four real datasets and all four simulations. The average difference in c-index on the validation sets (integrative - fast approximation) is 0.00, and the root mean-squared error is 0.02. The fully integrative model took approximately 11 minutes on 50 cores to fit and predict for the GBM $0^{th}$ cross-validation

set for Model 0, while the fast approximation took around 4 minutes and 1 core. For the LGG $0^{th}$ cross-validation set for Model 0, the fully integrative model took approximately 25 minutes, while the fast approximation took approximately 5 minutes.

The majority of the instances tested had the fast approximation perform equivalently to the the fully integrative model. However, while preparing the simulation studies, we observed some instances where the fully integrative model outperformed the fast approximation. We recommend using the fast approximation for exploratory analysis, and the fully integrative model for final analysis.

## SI-5 Concordance Index

The concordance index (c-index) was introduced by [15] as a non-parametric measure of survival time prediction accuracy. In the absence of censoring, the c-index is related to the Wilcoxon-Mann-Whitney U-statistic. We use the generalization of the c-index to account for ties introduced by [16]. With the following definition,

$$\tau_{(\tilde{t},\delta),(\hat{t},\hat{\delta})} = \frac{1}{N(N-1)} \sum_{n=1}^{N} \sum_{\substack{m=1 \\ m \neq n}}^{N} \left( \mathbf{I}(\tilde{t}_n \geq \tilde{t}_m)\delta_m - \mathbf{I}(\tilde{t}_n \leq \tilde{t}_m)\delta_n \right) \left( \mathbf{I}(\hat{t}_n \geq \hat{t}_m)\hat{\delta}_m - \mathbf{I}(\hat{t}_n \leq \hat{t}_m)\hat{\delta}_n \right)$$

the c-index that accounts for ties is,

$$c(\tilde{\mathbf{t}}, \boldsymbol{\delta}, \hat{\mathbf{t}}, \hat{\boldsymbol{\delta}} = \mathbb{1}) \quad = \quad \frac{1}{2} \left( \frac{\tau_{(\tilde{t},\delta),(\hat{t},\mathbb{1})}}{\tau_{(\tilde{t},\delta),(\hat{t},\delta)}} + 1 \right). \tag{SI-26}$$

This reduces to the standard c-index when there are no ties.

## SI-6 Cross-validation Strategy

The following outlines our cross-validation strategy:

1. Reserve 25% of the individuals (uniformly at random) as a test set. These individuals are not used in cross-validation. Divide the remaining 75% individuals (uniformly at random) into $n_{cv}$ cross-validation sets. Our results use $n_{cv} = 5$.

2. Select a learning model $\hat{M}$. This could be various $d_z$ for FA-EPH-C or various $L_1$ penalties for EPH-$L_1$.

3. For each $n_{cv}$ cross-validation set $v$, the learning set $l$ is the remaining $(n_{cv} - 1)/n_{cv}$ cross-validation sets. For each set $v$:

(a) Learn the parameters $\hat{\Theta}_l$ for learning model $\hat{M}$ from the learning set $l$. In this section, $\Theta$ is a collection of all the model parameters. If $\hat{M}$ is a FA-EPH-C model, use the approximate GEM algorithm outlined in Sec. SI-3.1. If $\hat{M}$ is a EPH-$L_1$ model, use the iterative least-squares procedure outlined in Sec. SI-1.1.

(b) Predict $\hat{t}_v = \mathbb{E}_{\hat{\Theta}_l}[t|\mathbf{x}_v^{(1)}, \ldots, \mathbf{x}_v^{(D)}]$ for all $N$ individuals in the validation set $v$ (for $\hat{M} \in$FA-EPH-C, Eqn. SI-25 or Eqn. SI-6 for $\hat{M} \in$ EPH-C-$L_1$).

(c) Calculate the c-index on the validation set given the learning set, $c(v)|l = c(\mathbf{t}_v, \mathbb{1}, \hat{\mathbf{t}}_v, \mathbb{1})|\hat{\Theta}_l$ (Eqn. SI-26).

4. Calculate the mean and standard deviation for the cross-validated c-index.

## SI-7 Model Selection Strategy

We use the following conservative selection criteria for "best" c-index among the tested models:

- Find the model with largest mean c-index. This is the"best" model, as long as the following condition is met: no other model's mean ± standard deviation is contained within with the largest mean ± standard deviation.
- If the condition is not met, find the model with the next largest mean c-index that has it's mean ± standard deviation contained within the largest mean ± standard deviation. This is the"best" model, as long as the following condition is met: no other model's mean ± standard deviation is contained within with the second largest mean ± standard deviation.
- And so on, for additional nesting.

## SI-8 Independent Censoring

The hazard function for the survival time, $h_T(t|\mathbf{x})$, after conditioning on time-independent covariates $\mathbf{x}$, is (page 3)[17],

$$h_T(t|\mathbf{x}) = \lim_{\epsilon \to 0} P(t \leq T < t + \epsilon | T \geq t, \mathbf{x}) = \frac{p(t|\mathbf{x})}{P(T > t|\mathbf{x})}. \tag{SI-27}$$

In this expression, the conditional probability density is $p(t|\mathbf{x})$, and the conditional survival function is $P(T > t|\mathbf{x})$. Let us also define $h_T(t|C > t, \mathbf{x})$,

$$h_T(t|C > t, \mathbf{x}) \quad = \quad \lim_{\epsilon \to 0} P(t \leq T < t + \epsilon | T \geq t, C \geq t, \mathbf{x}). \tag{SI-28}$$

The condition for independent censoring is, conditioned on time-independent covariates $\mathbf{x}$ (page 26-27)[17],

$$h_T(t|\mathbf{x}) \quad = \quad h_T(t|C > t, \mathbf{x}) \quad \text{whenever } P(\tilde{T} > t) > 0 \tag{SI-29}$$

For the probability distribution for FA-EPH-C, (Eqn. 15), the relevant conditional hazard functions are,

$$h_T(t|\mathbf{x}) \quad = \quad \frac{\int p_{\mathbf{w}_T}(t|\mathbf{z}) \left( \prod_{d=1}^{D} p_{\Theta^{(d)}}(\mathbf{x}^{(d)}|\mathbf{z}) \right) p(\mathbf{z}) d\mathbf{z}}{\int P_{\mathbf{w}_T}(T > t|\mathbf{z}) \left( \prod_{d=1}^{D} p_{\Theta^{(d)}}(\mathbf{x}^{(d)}|\mathbf{z}) \right) p(\mathbf{z}) d\mathbf{z}} \tag{SI-30}$$

$$h_T(t|C > t, \mathbf{x}) \quad = \quad \frac{\int p_{\mathbf{w}_T}(t|\mathbf{z}) P_{\mathbf{w}_C}(C > t|\mathbf{z}) \left( \prod_{d=1}^{D} p_{\Theta^{(d)}}(\mathbf{x}^{(d)}|\mathbf{z}) \right) p(\mathbf{z}) d\mathbf{z}}{\int P_{\mathbf{w}_T}(T > t|\mathbf{z}) P_{\mathbf{w}_C}(C > t|\mathbf{z}) \left( \prod_{d=1}^{D} p_{\Theta^{(d)}}(\mathbf{x}^{(d)}|\mathbf{z}) \right) p(\mathbf{z}) d\mathbf{z}}. \tag{SI-31}$$

It can be shown (Theorem 1)[18] the independent censoring condition (Eqn. SI-29) holds for FA-EPH-C if and essentially only if,

$$P_{\mathbf{w}_C}(C > t|\mathbf{z})^{-1} = \text{a constant} \quad \text{or} \quad h_T(t|\mathbf{z}) = \text{a constant}. \tag{SI-32}$$

In FA-EPH-C, this only occurs when the censoring distribution and/or the survival distribution is independent of the latent variables, and $\boldsymbol{\beta}_C = \mathbb{0}$ and/or $\boldsymbol{\beta}_T = \mathbb{0}$, respectively.

## SI-9 Additional Applications

### SI-9.1 Glioblastoma multiforme.

Glioblastoma multiforme (GBM) is the most common primary brain tumor in adults. Primary GBM arises *de novo* without progression from previously diagnosed LGG. Like LGG, primary GBM exhibits heterogeneity in molecular phenotype and survival response. Clinical, exome sequence, DNA copy number, DNA methylation, and messenger RNA expression data have been collected for many glioblastomas from adults.[19] Seventy-one primary glioblastomas have complete data including survival time, and 56 of the 71 patients are uncensored (79%). Table 1 contains the dimension of each type of covariate used in FA-EPH-C and EPH-C-$L_1$, and the data collection and data platforms are discussed in detail the original paper [19].

A cross-validated search for FA-EPH-C over latent dimensions $d_z = \{2, 3, 4, 5\}$ (Model numbers $0 - 3$) identifies latent dimension $d_z = 2$ (Model 0) (Fig. SI-6). The models for $d_z = \{3, 4, 5\}$ are eliminated because for at least one of the five CV sets, the EM algorithm approached a Heywood case [20]. A Heywood case has some components of an estimated $\Psi^{(d)}$ approach zero. One of the causes of Heywood cases is too many latent variables. A cross-validated search for sparsity parameters $\gamma \in \{5e3, 1e4, 1e5\}$ (Model numbers $4 - 6$) for EPH-C-$L_1$ identifies $\gamma = 1e5$ (Model 6) as the best penalty; however the performance of this model on the GBM data is still quite poor (Fig. SI-6). The search for PCA-EPH-C over latent dimensions $d_z = \{2, 3, 4, 5\}$ (Model numbers $7 - 10$) identifies dimension $d_z = 3$ (Model 8) as the best PCA-EPH-C model. For the GBM EPH-C gold-standard model (Model number 11), we include age (in years), a 5-class categorical variable capturing expression subtype (Classical, Mesenchymal, Neural, Proneural, G-CIMP), a binary variable capturing MGMT status ($\in \{methylated = 1, unmethylated = 0\}$), and a binary variable capturing $IDH1$ status ($\in \{WT = 1, R132H = 0\}$) as covariates. These covariates were identified as predictive of positive clinical outcomes.[19] The performance of the gold standard is worse than the other models considered in the CV model selection stage, despite the fact that the comparison is biased in favor of the gold-standard model. FA-EPH-C outperforms EPH-C-$L_1$, PCA-EPH-C, and gold standard models during CV.

At latent dimension $d_z = 2$ for FA-EPH-C, the latent projection of the training and validation data shows no appreciable clustering among the patients (Fig. SI-7). Fig. SI-8 shows that the latent projection found by FA-EPH-C $d_z = 2$ reveals some correlation with the $IDH1$ status, $MGMT$ status, and the expression subtype. The most apparent correlation is with $MGMT$ status.

The final test set c-index prediction accuracy for GBM was quite variable across the final models. The gold standard EPH-C model performed best, followed by PCA-EPH-C with $d_z = 3$, then FA-EPH-C with $d_z = 2$, and lastly the EPH-C-$L_1$ with $\gamma = 1e5$. The results are summarized in Table 2 .

### SI-9.2 Lung adenocarcinoma

Lung cancer is the leading cause of cancer-related mortality around the world, and lung adenocarcinoma (LUAD) is the most common type of lung cancer. Clinical, exome sequence, DNA copy number, messenger RNA expression, and micro RNA expression data have been collected for 172 LUAD tumors, with 72 patients uncensored (42%).[21] Table 1 contains

the dimension of each type of covariate used in FA-EPH-C and EPH-C-$L_1$. See the original paper for a comprehensive discussion of the data collection and platforms.[21]

A cross-validated search for FA-EPH-C over latent dimensions $d_z = \{2, 3, 4, 5\}$ (Model numbers $0-3$) identifies dimension $d_z = 3$ (Model 1) (Fig. SI-9). A cross-validated search for sparsity parameters $\gamma \in \{1e3, 1e4, 1e5\}$ (Model numbers $4-6$) for EPH-C-$L_1$ identifies $\gamma = 1e3$ (Model 4) as the best penalty (Fig. SI-9). The search for PCA-EPH-C over latent dimensions $d_z = \{2, 3, 4, 5\}$ (Model numbers $7 - 10$) identifies dimension $d_z = 3$ (Model 8) as the best PCA-EPH-C model. For the LUAD gold standard model (Model number 11), we take a 3-class categorical variable for expression subtype (Terminal respiratory unit (TRU), Proximal-proliferative (PP), Proximal-inflammatory (PI)), and a four-class categorical variable for the pathology N-stage (n0, n1, n2, nx). These covariates were associated with survival outcome, with the TRU subtype exhibiting superior outcomes.[21] This gold standard model is the best predictor of survival time at the CV stage (Fig. SI-9).

The latent projection of the training and validation data shows no apparent clustering among the patients at latent dimension $d_z = 2$ for FA-EPH-C (Fig. SI-13). The latent projection found by FA-EPH-C $d_z = 2$, Fig. SI-14, exhibits some correlation with N-stage and expression subtype. The most appreciable correlation is with the classical expression subtype.

The final test set c-index prediction accuracy for LUAD was quite poor across the final models, with the exception of PCA-EPH-C with $d_z = 3$. The gold standard EPH-C model performed next best, followed by the EPH-C-$L_1$ with $\gamma = 1e3$, and lastly by the FA-EPH-C with $d_z = 3$. The results are summarized in Table 2 .

### SI-9.3 Lung squamous cell carcinoma.

Lung squamous cell carcenoma (LUSC) is the second most common type of lung cancer. Clinical, exome sequence, DNA copy number, DNA methylation, and messenger RNA expression data have been collected for 104 LUSC tumors.[22] Table 1 contains the dimension of each type of covariate used in FA-EPH-C and EPH-C-$L_1$. Of the 104 patients, 47 are uncensored (45%). See the original paper for a comprehensive discussion of the data collection and platforms.[22]

A cross-validated search for FA-EPH-C over latent dimensions $d_z = \{2, 5, 10, 15\}$ (Model numbers $0 - 3$) identifies dimension $d_z = 15$ (Model 3) (Fig. SI-12). We search over larger latent dimensions for LUSC than for the other datasets because for LUSC, fitting the larger

dimensional models do not result in Heywood cases. A cross-validated search for sparsity parameters $\gamma \in \{1e3, 1e4, 1e5\}$ (Model numbers $4-6$) for EPH-C-$L_1$ identifies $\gamma = 1e3$ (Model 4) as the best penalty; this is the best performing model for CV (Fig. SI-12). The search for PCA-EPH-C over latent dimensions $d_z = \{2, 5, 10, 15\}$ (Model numbers $7-10$) identifies dimension $d_z = 2$ (Model 7) as the best PCA-EPH-C model, due to it's comparatively small variance. However, the performance is quite poor. The original paper does not perform a survival analysis on the cohort and does not identify specific covariates as predictive of positive clinical outcomes. [22] We follow later work [23] and use gender, smoking history, age at initial pathologic diagnosis, and tumor stage as covariates in our gold standard model (Model number 11). In particular, we use a binary variable for gender ($\in \{Female = 1, Male = 0\}$), a 4-class categorical variable for smoking history (Lifelong non-smoker, Current reformed smoker for $\leq 15$ years, Current reformed smoker for $> 15$ years, and Current smoker), and a 7-class categorical variable for pathology T-stage (t1, t1a, t1b, t2, t2a, t3, and t4). This gold standard model is a poor predictor of survival time (Fig. SI-12). FA-EPH-C outperforms the gold standard model during CV.

The latent projection of the training and validation data shows no apparent clustering among the patients at latent dimension $d_z = 2$ for FA-EPH-C (Fig. SI-13). The latent projection found by FA-EPH-C $d_z = 2$, Fig. SI-14, exhibits some correlation with the smoking status, gender, and T-stage. The most appreciable correlation is with T-stage.

The final test set c-index prediction accuracy for LUSC was quite poor across all of the final models. The FA-EPH-C with $d_z = 15$ model performed best, followed by the EPH-C-$L_1$ with $\gamma = 1e3$ and PCA-EPH-C with $d_z = 2$, and last by the gold standard EPH-C model. The results are summarized in Table 2 .

## SI-10 Additional Figures

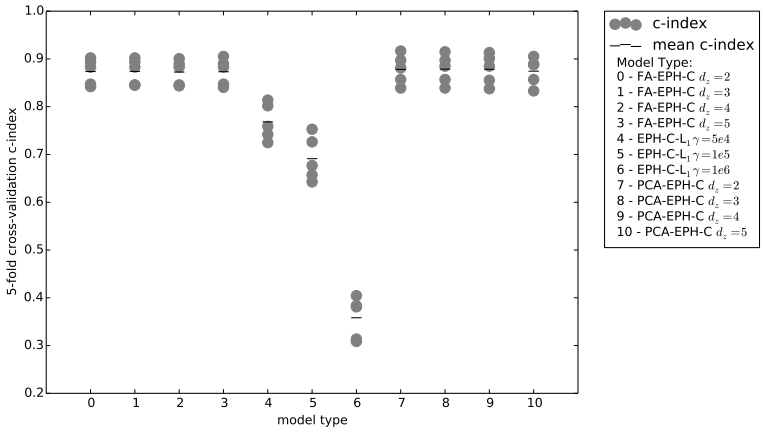



Figure SI-1: Results for the SIM4 5-fold CV latent dimension search for FA-EPH-C and comparison to the EPH-C-$L_1$ model. Model types are as follows. Models $0-3$ are FA-EPH-C and have, in order, $d_z = \{2,3,4,5\}$. Models $4-6$ are EPH-C-$L_1$ and have, in order, $\gamma = \{5e4, 1e5, 1e6\}$, which selects an average of $\{5,5,5\}$ relevant covariates.

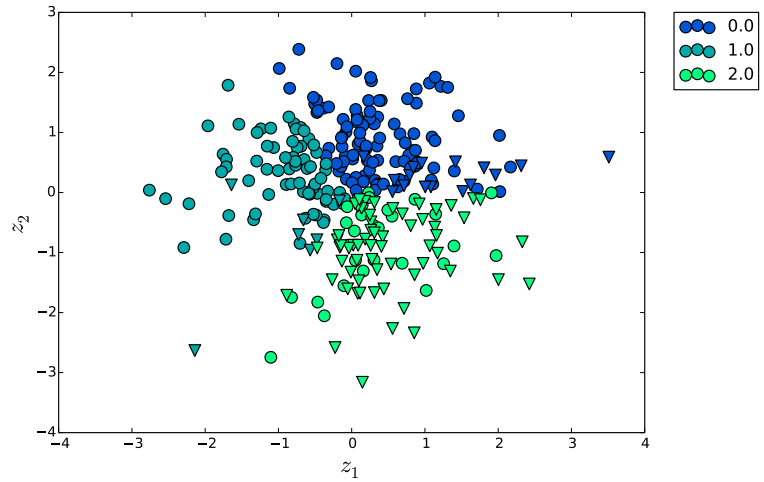



Figure SI-2: All of the SIM4 samples in the true latent space $\mathbf{z}$.

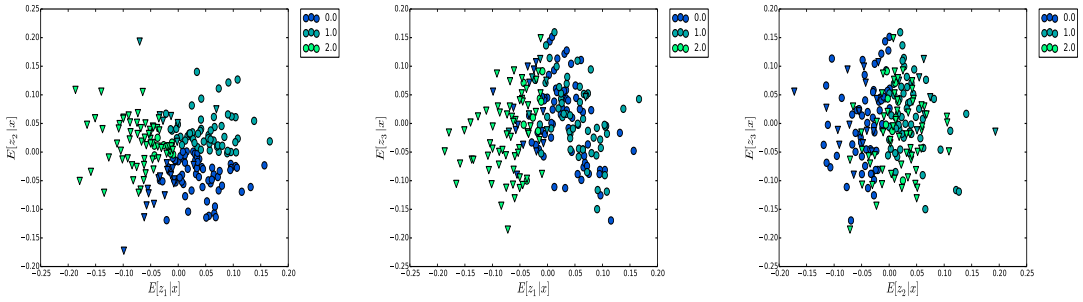



Figure SI-3: SIM4 latent projections $\mathbb{E}[\mathbf{z}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}]$ of the training cohort for the FA-EPH-C $d_z = 3$ model, colored by group membership. Circles represent uncensored observations, and triangles represent censored observations.

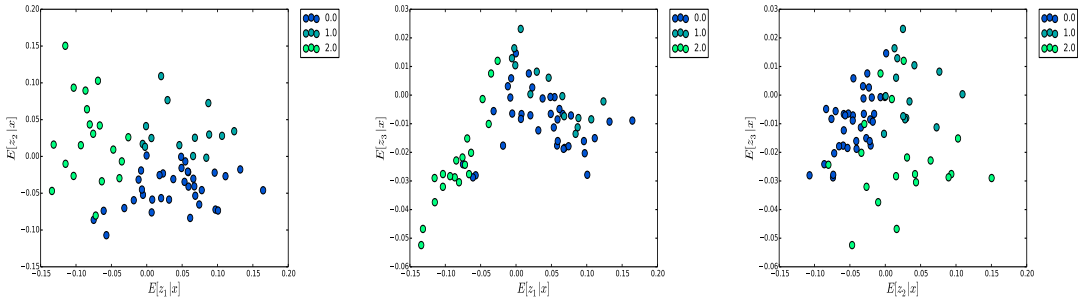
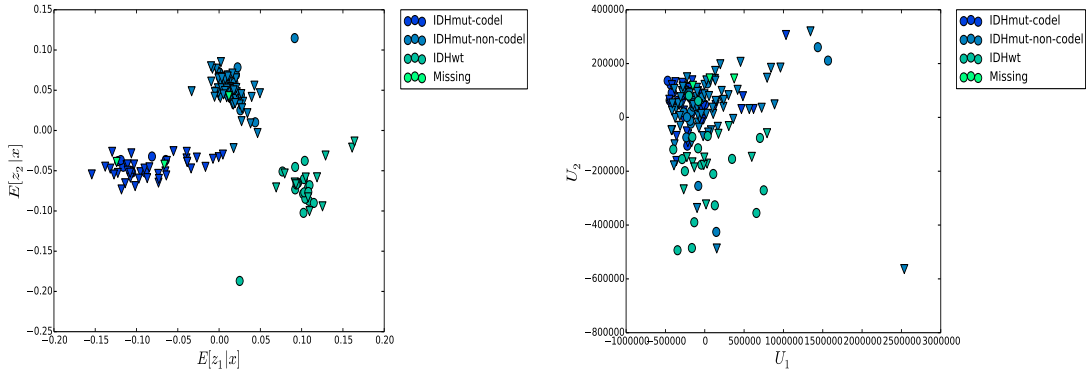



Figure SI-4: SIM4 latent projections $\mathbb{E}[\mathbf{z}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}]$ of the test cohort for the FA-EPH-C $d_z = 3$ model, colored by group membership. Circles represent uncensored observations, and triangles represent censored observations.




(a) $IDH - 1p/19q$ Status, Training Set, FA-EPH-C (b) $IDH - 1p/19q$ Status, Training Set, PCA

Figure SI-5: LGG latent projections $\mathbb{E}[\mathbf{z}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}]$ of the $0^{th}$ cross-validation training cohort for the FA-EPH-C $d_z = 2$ model, and the $d_z = 2$ PCA projection of the same $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}$ data. Circles represent uncensored observations, and triangles represent censored observations.

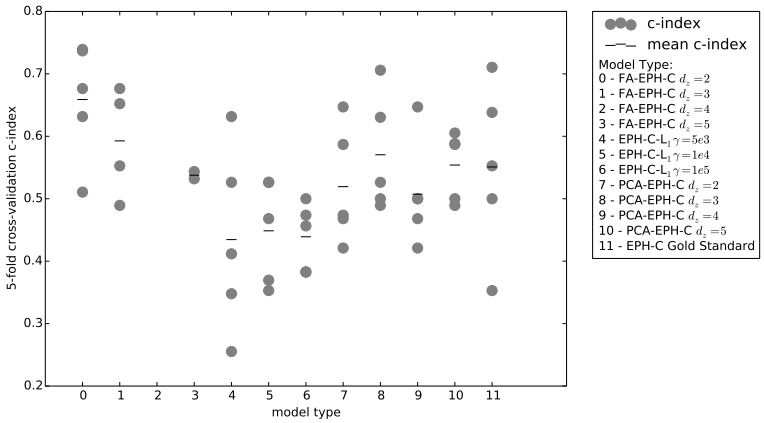
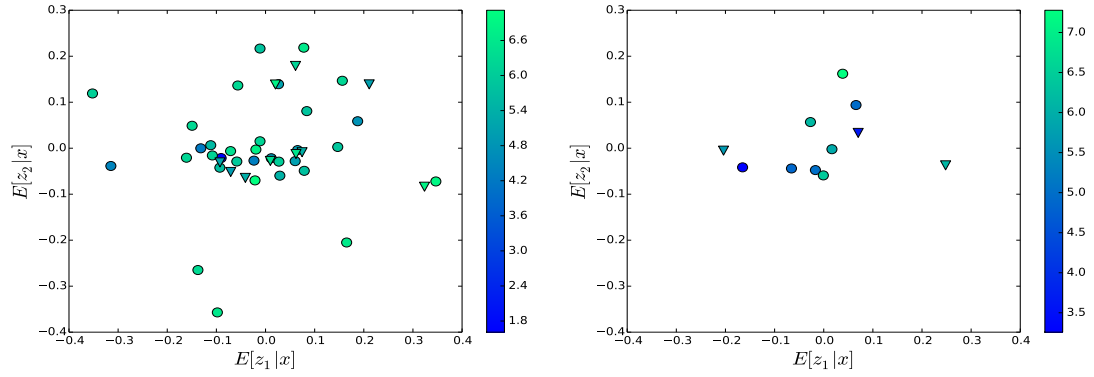
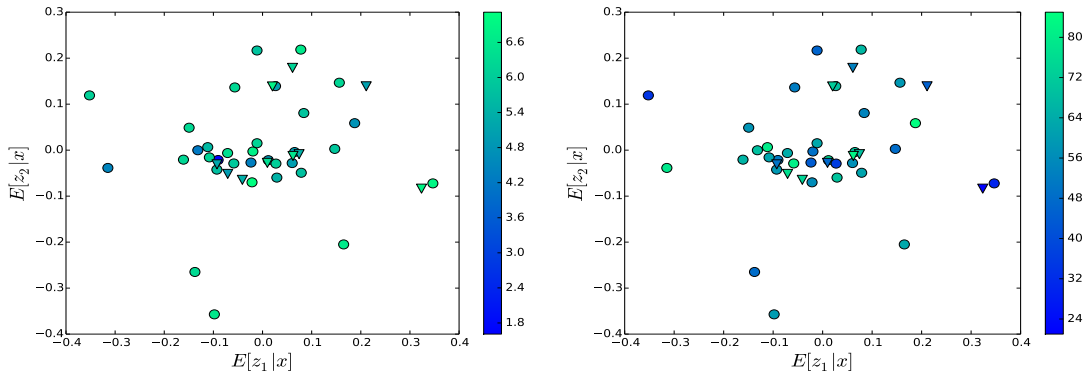



Figure SI-6: Results for the GBM 5-fold cross-validation latent dimension search for FA-EPH-C, comparison to EPH-C-$L_1$, and gold standard EPH-C. Model types are as follows. Models $0-3$ are FA-EPH-C and have, in order, $d_z = \{2, 3, 4, 5\}$. Models $1, 2, 3$ are eliminated from the model selection search because on at least 1 out of the 5 cross-validation groups, the EM algorithm approaches a Heywood case. Models $4-6$ are EPH-C-$L_1$ and have, in order, $\gamma = \{5e3, 1e4, 1e5\}$, which selects on average $\{14, 9, 2\}$ relevant covariates. Models $7-10$ are PCA-EPH-C with $d_z = \{2, 3, 4, 5\}$. Model 11 is the gold standard EPH-C model.




(a) Log Event Time, in Log days, Training Set 0 (b) Log Event Time, in Log days, Validation Set 0

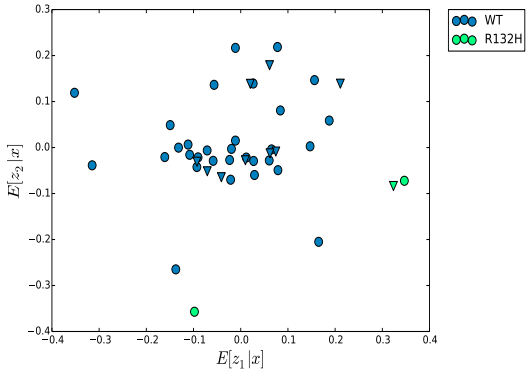
Figure SI-7: GBM latent projections $\mathbb{E}[\mathbf{z}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}]$ of the $0^{th}$ cross-validation training and validation cohort for the FA-EPH-C $d_z = 2$ model. Circles represent uncensored observations, and triangles represent censored observations.



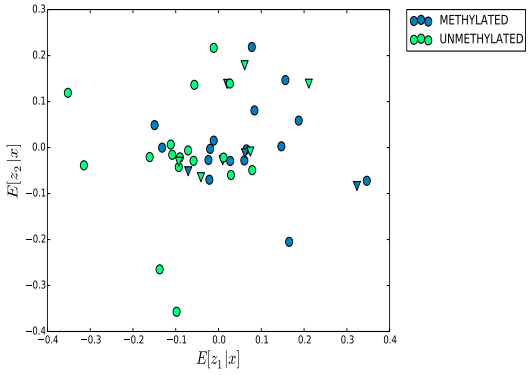

(a) Log Event Time, in Log days



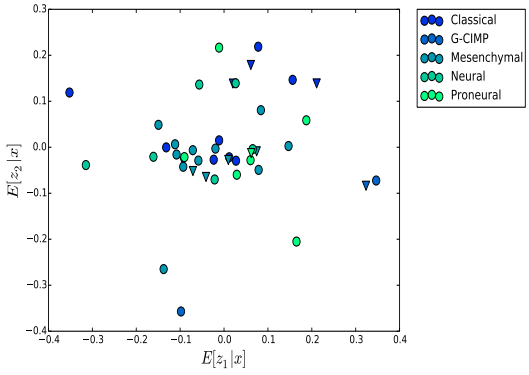

(b) Age, in years




(c) IDH1 Status




(d) MGMT Status




(e) Expression Subtype

Figure SI-8: GBM latent projections $\mathbb{E}[\mathbf{z}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}]$ of the $0^{th}$ cross-validation training cohort for the FA-EPH-C $d_z = 2$ model. Circles represent uncensored observations, and triangles represent censored observations.

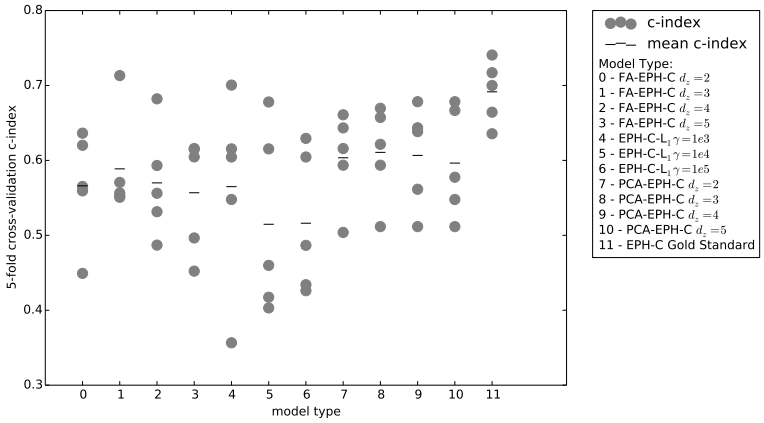
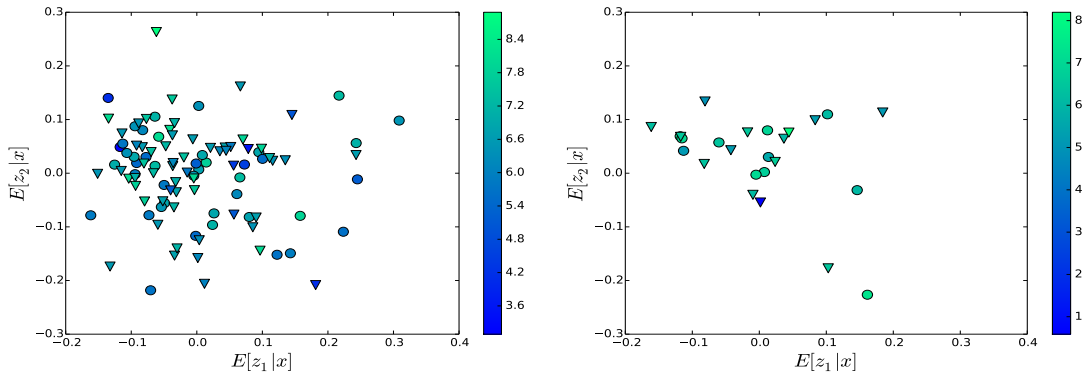
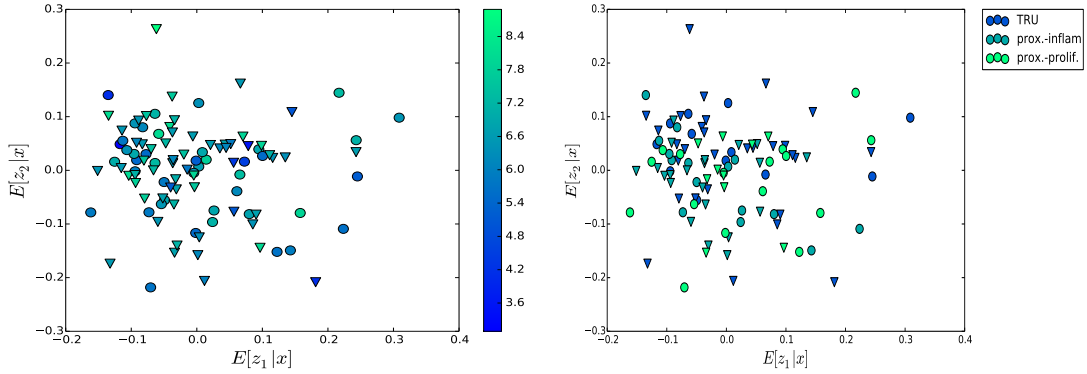



Figure SI-9: Results for the LUAD 5-fold cross-validation latent dimension search for FA-EPH-C, comparison to EPH-C-$L_1$, and gold standard EPH-C. Model types are as follows. Models $0-4$ are FA-EPH-C and have, in order, $d_z = \{2, 3, 4, 5\}$. Models $4-6$ are EPH-C-$L_1$ and have, in order, $\gamma = \{1e3, 1e4, 1e5\}$, which selects on average $\{39, 12, 3\}$ relevant covariates. Models $7-10$ are PCA-EPH-C with $d_z = \{2, 3, 4, 5\}$. Model 11 is the gold standard EPH-C model.




(a) Log Event Time, in Log days, Training Set 0 (b) Log Event Time, in Log days, Validation Set 0

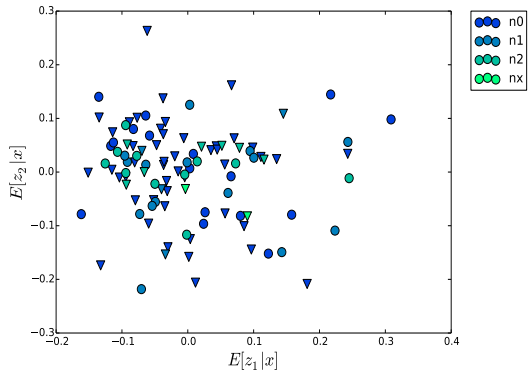
Figure SI-10: LUAD latent projections $\mathbb{E}[\mathbf{z}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}]$ of the $0^{th}$ cross-validation training and validation cohort for the FA-EPH-C $d_z = 2$ model. Circles represent uncensored observations, and triangles represent censored observations.




(a) Log Event Time, in Log days




(b) Expression Subtype




(c) Pathology N-stage

Figure SI-11: LUAD latent projections $\mathbb{E}[\mathbf{z}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}]$ of the $0^{th}$ cross-validation training cohort for the FA-EPH-C $d_z = 2$ model. Circles represent uncensored observations, and triangles represent censored observations.

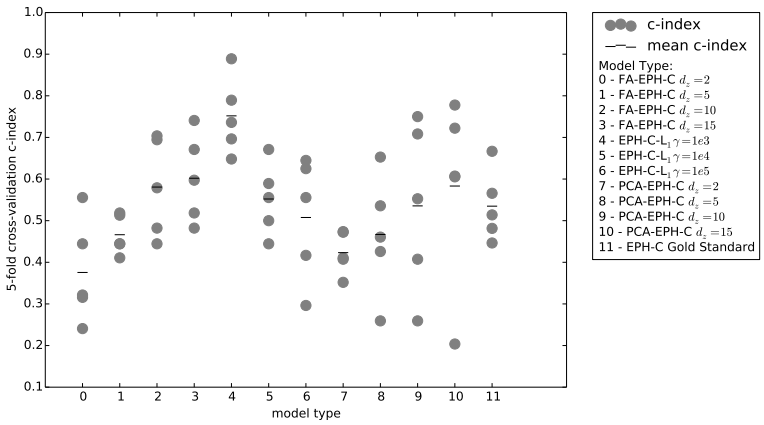
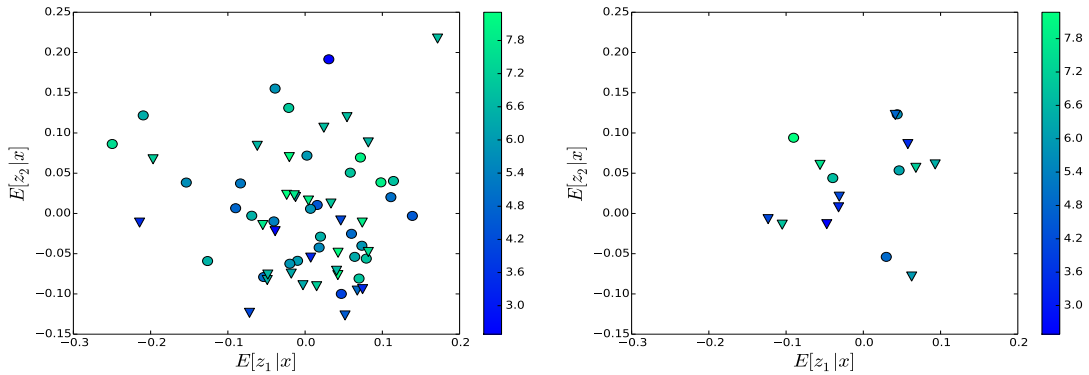
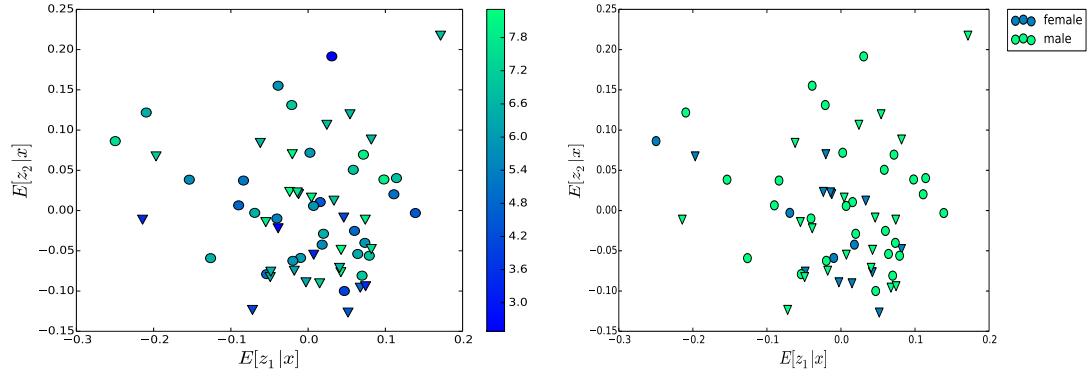



Figure SI-12: Results for the LUSC 5-fold cross-validation latent dimension search for FA-EPH-C, comparison to EPH-C-$L_1$, and gold standard EPH-C. Model types are as follows. Models $0-3$ are FA-EPH-C and have, in order, $d_z = \{2, 5, 10, 15\}$. Models $4-6$ are EPH-C-$L_1$ and have, in order, $\gamma = \{1e3, 1e4, 1e5\}$, which selects an average of $\{29, 11, 6\}$ relevant covariates. Models $7-10$ are PCA-EPH-C with $d_z = \{2, 5, 10, 15\}$. Model 11 is the gold standard EPH-C model.




(a) Log Event Time, in Log days, Training Set 0 (b) Log Event Time, in Log days, Validation Set 0

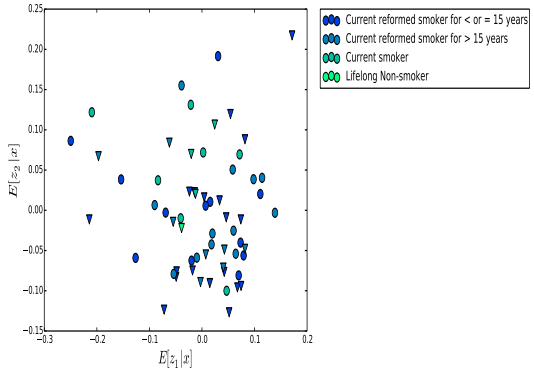
Figure SI-13: LUSC latent projections $\mathbb{E}[\mathbf{z}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}]$ of the $0^{th}$ cross-validation training and validation cohort for the FA-EPH-C $d_z = 2$ model. Circles represent uncensored observations, and triangles represent censored observations.



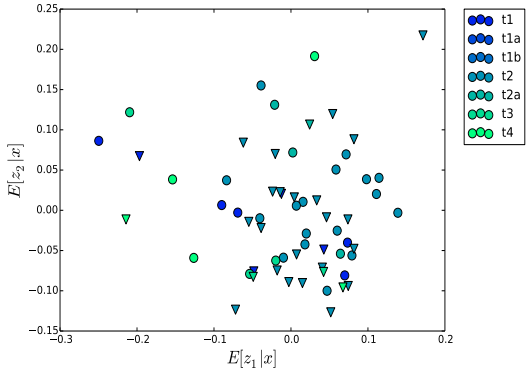

(a) Log Event Time, in Log days




(b) Gender




(c) Smoking




(d) Pathology T stage

Figure SI-14: LUSC latent projections $\mathbb{E}[\mathbf{z}|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(D)}]$ of the $0^{th}$ cross-validation training cohort for the FA-EPH-C $d_z = 2$ model. Circles represent uncensored observations, and triangles represent censored observations.

# References


[1] Tibshirani R.. The lasso method for variable selection in the Cox model. *Statistics in Medicine.* 1997;16(4):385–395.

[2] Efron Bradley, Hastie Trevor, Johnstone Iain, Tibshirani Robert. Least angle regression. *The Annals of Statistics.* 2004;32(2):407–499.

[3] Dempster A. P., Laird N. M., Rubin D. B.. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological).* 1977;39(1):1–38.

[4] Rubin Donald B., Thayer Dorothy T.. EM algorithms for ML factor analysis. *Psychometrika.* 1982;47(1):69–76.

[5] Jaakkola Tommi S., Jordan Michael I.. A variational approach to Bayesian logistic regression models and their extensions. In: ; 1997.

[6] Tipping Michael. Probabilistic Visualisation of High-dimensional Binary Data. *Advances in Neural Information Processing Systems 11.* 1999;:592–598.

[7] Meng Xiao-Li, Rubin Donald B.. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika.* 1993;80(2):267–278.

[8] Bouchard Guillaume. *Efficient bounds for the softmax function, applications to inference in hybrid models, Presentation at the Workshop for Approximate Bayesian Inference in Continuous/Hybrid Systems at NIPS-07.* 2007.

[9] Tipping Michael E., Bishop Christopher M.. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 1999;61(3):611–622.

[10] Metropolis Nicholas, Rosenbluth Arianna W., Rosenbluth Marshall N., Teller Augusta H., Teller Edward. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics.* 1953;21(6):1087–1092.

[11] Hastings W. K.. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika.* 1970;57(1):97–109.

[12] Wei Greg C. G., Tanner Martin A.. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association.* 1990;85(411):699–704.

[13] Gelman Andrew, Carlin John B., Stern Hal S., Dunson David B., Vehtari Aki, Rubin Donald B.. *Bayesian Data Analysis, Third Edition.* Boca Raton: Chapman and Hall/CRC; 3 edition ed.2013.

[14] Wu C. F. Jeff. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics.* 1983;11(1):95–103.

[15] Harrell , Robert M. Califf , David B. Pryor , Kerry L. Lee , Robert A. Rosati . Evaluating the yield of medical tests. *JAMA.* 1982;247(18):2543–2546.

[16] Kang Le, Chen Weijie, Petrick Nicholas A., Gallas Brandon D.. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. *Statistics in Medicine.* 2015;34(4):685–703.

[17] Fleming Thomas R., Harrington David P.. *Counting processes and survival analysis.* Wiley series in probability and statisticsHoboken, N.J: Wiley-Interscience; 2005.

[18] Wijsman Robert A.. A Useful Inequality on Ratios of Integrals, With Application to Maximum Likelihood Estimation. *Journal of the American Statistical Association.* 1985;80(390):472–475.

[19] Brennan Cameron W., Verhaak Roel G.W., McKenna Aaron, et al. The Somatic Genomic Landscape of Glioblastoma. *Cell.* 2013;155(2):462–477.

[20] Heywood H. B.. On Finite Sequences of Real Numbers. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character.* 1931;134(824):486–501.

[21] The Cancer Genome Atlas Research Network . Comprehensive molecular profiling of lung adenocarcinoma. *Nature.* 2014;511(7511):543–550.

[22] The Cancer Genome Atlas Research Network . Comprehensive genomic characterization of squamous cell lung cancers. *Nature.* 2012;489(7417):519–525.

[23] Jiang Liyan, Zhu Wei, Streicher Katie, et al. Increased IR-A/IR-B ratio in non-small cell lung cancers associates with lower epithelial-mesenchymal transition signature and longer survival in squamous cell lung carcinoma. *BMC cancer.* 2014;14:131.