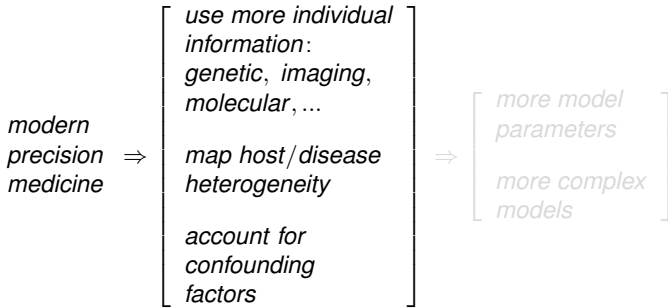
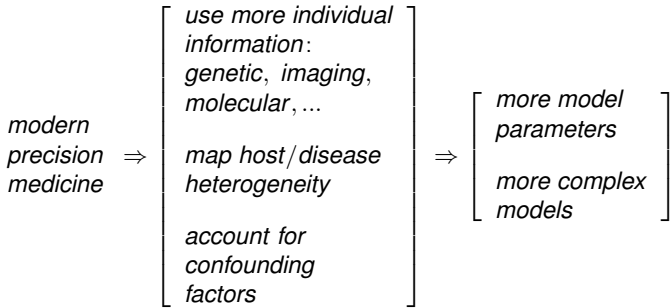
Overfitting correction in multivariate survival analysis

ACC Coolen
King's College London and Saddle Point Science
Belfast, Sept 4th 2019

modern precision medicine $\Rightarrow$ [*use more individual information: genetic, imaging, molecular, ...*; *map host/disease heterogeneity*; *account for confounding factors*] $\Rightarrow$ [*more model parameters*; *more complex models*]


modern
precision
medicine
⇒
use more individual information: genetic, imaging, molecular, ...
map host/disease heterogeneity
account for confounding factors
⇒
more model parameters
more complex models


modern
precision
medicine
⇒
use more individual information: genetic, imaging, molecular, ...
map host/disease heterogeneity
account for confounding factors
⇒
more model parameters
more complex models

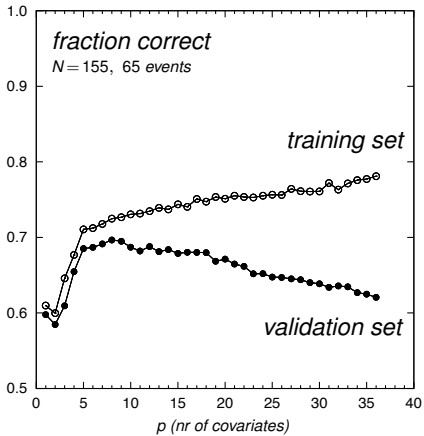
modern precision medicine $\Rightarrow$ [*use more individual information*: *genetic*, *imaging*, *molecular*, ... ; *map host/disease heterogeneity* ; *account for confounding factors*] $\Rightarrow$ [*more model parameters* ; *more complex models*] $\Rightarrow$ *overfitting* 

Phenomenology of overfitting

deteriorating outcome prediction performance on unseen data ...

multivariate Cox regression:

predict whether event before or after a cutoff time point



primitive rule of thumb:
$p_{\max} \sim \#events/10$

- too optimistic?
- indep of association strengths?
- indep of covariate correlations?

Phenomenology of overfitting

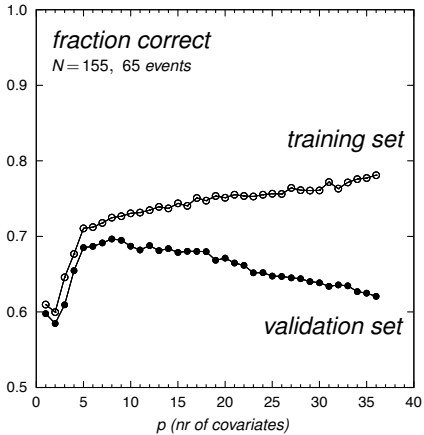
deteriorating outcome prediction performance on unseen data ...

multivariate Cox regression:

predict whether event before or after a cutoff time point

primitive rule of thumb:
$p_{\max} \sim \#events/10$

- too optimistic?
- indep of association strengths?
- indep of covariate correlations?



false positive associations ...

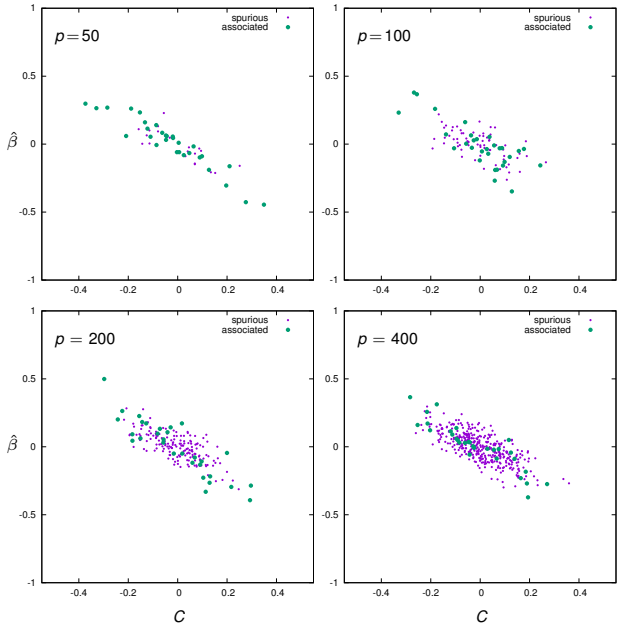
$N = 100$

p covariates:

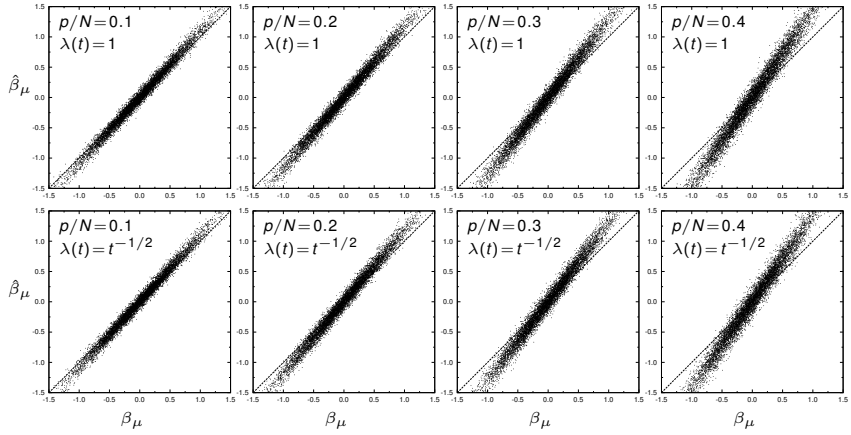
30 true associations
$p - 30$ spurious ones

C: Pearson correlation between covariates and event time,

$\hat{\beta}$: univariate Cox parameters



bias in inferred association parameters ...

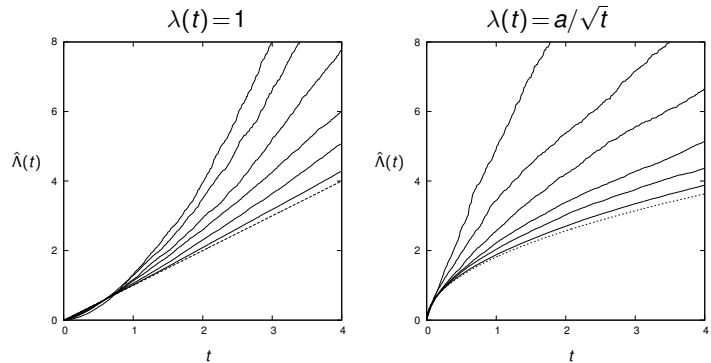


β_μ: true associations
$\hat{\beta}_\mu$: multivariate regression

synthetic survival data, generated from Cox model with $N = 400$

bias in inferred base hazard rates ...

$\hat{\Lambda}(t) = \int_0^t \mathrm{d}t' \; \lambda(t')$
$p/N = 0.05 \to 0.55$



synthetic survival data,
generated from Cox model with $N = 400$

Failure of our low-dimensional intuition

low-dim ML/MAP regime: $N \to \infty$, p fixed
high-dim regime: $N, p \to \infty$, p/N finite

- p-values are blind to overfitting ...
- hyperparameters of priors (regularizers) must be p-dependent ...
- for sufficiently large N:
  regression possible even when $p > N$ ...

notation:

$$\textit{data}: \quad \mathcal{D} = \{(\boldsymbol{z}_1, t_1), \ldots, (\boldsymbol{z}_N, t_N)\}, \qquad \begin{matrix} \boldsymbol{z}_i \in \mathbb{R}^p & \textit{covariates} \\ t_i > 0 & \textit{event time} \end{matrix}$$

$$\textit{Cox model}: \quad p(t|\boldsymbol{z}, \beta, \lambda) = -\frac{\mathrm{d}}{\mathrm{d}t} \exp[-\mathrm{e}^{\beta \cdot \boldsymbol{z}} \Lambda(t)]$$

$$\textit{ML inference}: \quad (\hat{\beta}, \hat{\lambda}) = \operatorname{argmax}_{\beta,\lambda} \Big\{ \sum_{i=1}^{N} \log p(t_i|\boldsymbol{z}_i, \beta, \lambda) \Big\}$$

$$\textit{MAP inference}: \quad (\hat{\beta}, \hat{\lambda}) = \operatorname{argmax}_{\beta,\lambda} \Big\{ \sum_{i=1}^{N} \log p(t_i|\boldsymbol{z}_i, \beta, \lambda) + \log \overbrace{p(\beta)}^{\textit{prior}} \Big\}$$

Failure of our low-dimensional intuition

low-dim ML/MAP regime: $N \to \infty$, p fixed
high-dim regime: $N, p \to \infty$, p/N finite

- p-values are blind to overfitting ...
- hyperparameters of priors (regularizers) must be p-dependent ...
- for sufficiently large N:
  regression possible even when $p > N$ ...

notation:

$$\textit{data}: \quad \mathcal{D} = \{(\boldsymbol{z}_1, t_1), \ldots, (\boldsymbol{z}_N, t_N)\}, \quad \begin{matrix} \boldsymbol{z}_i \in \mathbb{R}^p & \textit{covariates} \\ t_i > 0 & \textit{event time} \end{matrix}$$

$$\textit{Cox model}: \quad p(t|\boldsymbol{z}, \boldsymbol{\beta}, \lambda) = -\frac{\mathrm{d}}{\mathrm{d}t} \exp[-\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}} \Lambda(t)]$$

$$\textit{ML inference}: \quad (\hat{\boldsymbol{\beta}}, \hat{\lambda}) = \mathrm{argmax}_{\boldsymbol{\beta},\lambda} \Big\{ \sum_{i=1}^{N} \log p(t_i|\boldsymbol{z}_i, \boldsymbol{\beta}, \lambda) \Big\}$$

$$\textit{MAP inference}: \quad (\hat{\boldsymbol{\beta}}, \hat{\lambda}) = \mathrm{argmax}_{\boldsymbol{\beta},\lambda} \Big\{ \sum_{i=1}^{N} \log p(t_i|\boldsymbol{z}_i, \boldsymbol{\beta}, \lambda) + \log \overbrace{p(\boldsymbol{\beta})}^{\textit{prior}} \Big\}$$

scaling of hyperparameters

in regularised Cox regression

$$(\hat{\boldsymbol{\beta}}, \hat{\lambda}) = \text{argmax}_{\boldsymbol{\beta},\lambda} \Big\{ \sum_{i=1}^{N} \log p(t_i|\boldsymbol{z}_i, \boldsymbol{\beta}, \lambda) + \log p(\boldsymbol{\beta}) \Big\}$$

$$p(t|\boldsymbol{z}, \boldsymbol{\beta}, \lambda) = -\frac{\mathrm{d}}{\mathrm{d}t} \exp[-\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}} \Lambda(t)], \qquad \boldsymbol{\beta}\cdot\boldsymbol{z} = \sum_{\mu=1}^{p} \beta_\mu z_\mu$$

e.g. $\quad p(\boldsymbol{\beta}) \propto \mathrm{e}^{-\sum_{\mu=1}^{p} |\beta_\mu/\sigma|}, \qquad p(\boldsymbol{\beta}) \propto \mathrm{e}^{-\frac{1}{2}\sum_{\mu=1}^{p} (\beta_\mu/\sigma)^2}, \qquad \sigma:$ *hyperpar*

claim: $\quad \sigma = \mathcal{O}(p^{-\frac{1}{2}})$ as $p \to \infty$

- general theory
- simple scaling argument:

  $\beta_\mu = \mathcal{O}(1)$: $\quad \sum_{\mu=1}^{p} \beta_\mu z_\mu^i = \mathcal{O}(\sqrt{p}) \quad \Rightarrow \quad t_i \in \{0, \infty\}$

  $\beta_\mu = \mathcal{O}(p^{-1/2})$: $\quad \sum_{\mu=1}^{p} \beta_\mu z_\mu^i = \mathcal{O}(1) \quad \Rightarrow \quad t_i$ finite

finite event times = prior knowledge!

scaling of hyperparameters

in regularised Cox regression

$$(\hat{\boldsymbol{\beta}}, \hat{\lambda}) = \operatorname{argmax}_{\boldsymbol{\beta},\lambda} \Big\{ \sum_{i=1}^{N} \log p(t_i|\boldsymbol{z}_i, \boldsymbol{\beta}, \lambda) + \log p(\boldsymbol{\beta}) \Big\}$$

$$p(t|\boldsymbol{z}, \boldsymbol{\beta}, \lambda) = -\frac{\mathrm{d}}{\mathrm{d}t} \exp[-\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}} \Lambda(t)], \qquad \boldsymbol{\beta}\cdot\boldsymbol{z} = \sum_{\mu=1}^{p} \beta_\mu z_\mu$$

e.g. $\quad p(\boldsymbol{\beta}) \propto \mathrm{e}^{-\sum_{\mu=1}^{p} |\beta_\mu/\sigma|}, \qquad p(\boldsymbol{\beta}) \propto \mathrm{e}^{-\frac{1}{2}\sum_{\mu=1}^{p} (\beta_\mu/\sigma)^2}, \qquad \sigma:$ *hyperpar*

claim: $\quad \sigma = \mathcal{O}(p^{-\frac{1}{2}})$ as $p \to \infty$

- general theory
- simple scaling argument:

  $\beta_\mu = \mathcal{O}(1)$: $\quad \sum_{\mu=1}^{p} \beta_\mu z_\mu^i = \mathcal{O}(\sqrt{p}) \quad \Rightarrow \quad t_i \in \{0, \infty\}$

  $\beta_\mu = \mathcal{O}(p^{-1/2})$: $\quad \sum_{\mu=1}^{p} \beta_\mu z_\mu^i = \mathcal{O}(1) \quad \Rightarrow \quad t_i$ finite

finite event times = prior knowledge!

scaling of hyperparameters

in regularised Cox regression

$$(\hat{\boldsymbol{\beta}}, \hat{\lambda}) = \operatorname{argmax}_{\boldsymbol{\beta},\lambda} \Big\{ \sum_{i=1}^{N} \log p(t_i|\boldsymbol{z}_i, \boldsymbol{\beta}, \lambda) + \log p(\boldsymbol{\beta}) \Big\}$$

$$p(t|\boldsymbol{z}, \boldsymbol{\beta}, \lambda) = -\frac{\mathrm{d}}{\mathrm{d}t} \exp[-\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}} \Lambda(t)], \qquad \boldsymbol{\beta}\cdot\boldsymbol{z} = \sum_{\mu=1}^{p} \beta_\mu z_\mu$$

e.g. $\quad p(\boldsymbol{\beta}) \propto \mathrm{e}^{-\sum_{\mu=1}^{p} |\beta_\mu/\sigma|}, \qquad p(\boldsymbol{\beta}) \propto \mathrm{e}^{-\frac{1}{2}\sum_{\mu=1}^{p} (\beta_\mu/\sigma)^2}, \qquad \sigma:$ *hyperpar*

claim: $\quad \sigma = \mathcal{O}(p^{-\frac{1}{2}})$ as $p \to \infty$

- general theory
- simple scaling argument:

  $\beta_\mu = \mathcal{O}(1)$: $\quad \sum_{\mu=1}^{p} \beta_\mu z_\mu^i = \mathcal{O}(\sqrt{p}) \quad \Rightarrow \quad t_i \in \{0, \infty\}$

  $\beta_\mu = \mathcal{O}(p^{-1/2})$: $\quad \sum_{\mu=1}^{p} \beta_\mu z_\mu^i = \mathcal{O}(1) \quad \Rightarrow \quad t_i$ finite

finite event times = prior knowledge!

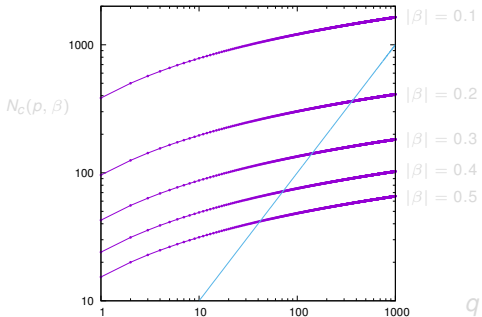
regression with $p > N$ in principle possible

provided N is large enough

- $N \uparrow$: prob of false positive associations $\downarrow$
- $p \uparrow$: prob of false positive associations $\uparrow$

uncorr covars: $N > N_c(p, \beta)$: *prob of finding one or more spurious univariate associations of strength* $\geq |\beta|$ *is less than* q

$$N_c(p, \beta) = \frac{2}{\beta^2}\left[\mathrm{Erf}^{-1}\left(e^{\frac{1}{p}\log(1-q)}\right)\right]^2$$



$q = 0.05$

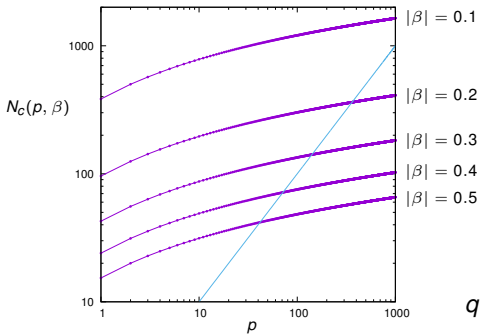
regression with $p > N$ in principle possible

provided N is large enough

- $N \uparrow$: prob of false positive associations $\downarrow$
- $p \uparrow$: prob of false positive associations $\uparrow$

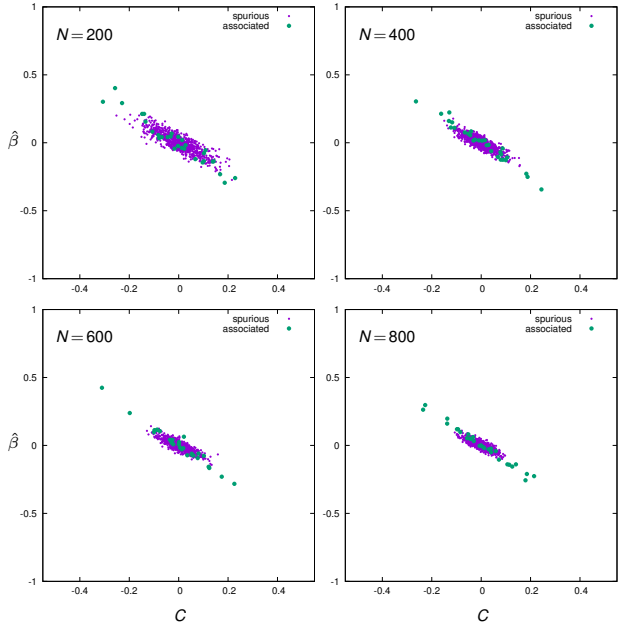
uncorr covars: $N > N_c(p, \beta)$: *prob of finding one or more spurious univariate associations of strength* $\geq |\beta|$ *is less than* q

$$N_c(p, \beta) = \frac{2}{\beta^2}\left[\mathrm{Erf}^{-1}\left(e^{\frac{1}{p}\log(1-q)}\right)\right]^2$$



$q = 0.05$

synthetic survival data,
with $p = 800$



Intuition for the problem

- information-theoretic interpretation of ML regression

  assumed model: $p_{\boldsymbol{\theta}}$

$$\begin{aligned} \boldsymbol{\theta}_{\mathrm{ML}} &= \mathrm{argmax}_{\boldsymbol{\theta}}\ p(\mathcal{D}|\boldsymbol{\theta}) \\ &= \mathrm{argmin}_{\boldsymbol{\theta}}\ D(\hat{p}||p_{\boldsymbol{\theta}}) \end{aligned}$$

$$\hat{p}(t,\boldsymbol{z}) = \frac{1}{N}\sum_{i=1}^{N}\delta(t-t_i)\delta(\boldsymbol{z}-\boldsymbol{z}_i) \qquad (\textit{empirical distribution})$$

$$D(\hat{p}||p_{\boldsymbol{\theta}}) = \int \mathrm{d}t\mathrm{d}\boldsymbol{z}\ \hat{p}(t,\boldsymbol{z}) \log\Big[\frac{\hat{p}(t|\boldsymbol{z})}{p(t|\boldsymbol{z},\boldsymbol{\theta})}\Big] \qquad (KL-\textit{distance})$$

- so ML regression pushes $p(t|\boldsymbol{z},\theta)$
  as close as possible towards $\hat{p}(t|\boldsymbol{z})$

  true pars: $\theta^\star$

  - fixed p, $N\to\infty$: $\quad \hat{p}(t,\boldsymbol{z}) = p(t,\boldsymbol{z}|\theta^\star), \quad$ so $\ \boldsymbol{\theta}_{\mathrm{ML}} = \theta^\star$ ✓
  - $p=\mathcal{O}(N)$, $N\to\infty$: $\quad \hat{p}(t,\boldsymbol{z}) \neq p(t,\boldsymbol{z}|\theta^\star), \quad$ so $\ \boldsymbol{\theta}_{\mathrm{ML}} \neq \theta^\star$ ✗

Intuition for the problem

- information-theoretic interpretation of ML regression

  assumed model: $p_{\boldsymbol{\theta}}$

$$\begin{aligned} \boldsymbol{\theta}_{\mathrm{ML}} &= \mathrm{argmax}_{\boldsymbol{\theta}}\ p(\mathcal{D}|\boldsymbol{\theta}) \\ &= \mathrm{argmin}_{\boldsymbol{\theta}}\ D(\hat{p}||p_{\boldsymbol{\theta}}) \end{aligned}$$

$$\hat{p}(t,\boldsymbol{z}) = \frac{1}{N}\sum_{i=1}^{N}\delta(t-t_i)\delta(\boldsymbol{z}-\boldsymbol{z}_i) \qquad (\textit{empirical distribution})$$

$$D(\hat{p}||p_{\boldsymbol{\theta}}) = \int \mathrm{d}t\mathrm{d}\boldsymbol{z}\ \hat{p}(t,\boldsymbol{z}) \log\Big[\frac{\hat{p}(t|\boldsymbol{z})}{p(t|\boldsymbol{z},\boldsymbol{\theta})}\Big] \quad (KL-\textit{distance})$$

- so ML regression pushes $p(t|\boldsymbol{z},\boldsymbol{\theta})$
  as close as possible towards $\hat{p}(t|\boldsymbol{z})$

  true pars: $\theta^\star$

  - fixed p, $N\to\infty$: $\quad \hat{p}(t,\boldsymbol{z}) = p(t,\boldsymbol{z}|\theta^\star)$, $\quad$ so $\theta_{\mathrm{ML}} = \theta^\star$ ✓
  - $p=\mathcal{O}(N)$, $N\to\infty$: $\quad \hat{p}(t,\boldsymbol{z}) \neq p(t,\boldsymbol{z}|\theta^\star)$, $\quad$ so $\theta_{\mathrm{ML}} \neq \theta^\star$ ✗

Intuition for the problem

- information-theoretic interpretation of ML regression

  assumed model: $p_{\boldsymbol{\theta}}$

$$\begin{aligned} \boldsymbol{\theta}_{\mathrm{ML}} &= \mathrm{argmax}_{\boldsymbol{\theta}}\; p(\mathcal{D}|\boldsymbol{\theta}) \\ &= \mathrm{argmin}_{\boldsymbol{\theta}}\; D(\hat{p}||p_{\boldsymbol{\theta}}) \end{aligned}$$

$$\hat{p}(t,\boldsymbol{z}) = \frac{1}{N}\sum_{i=1}^{N}\delta(t-t_i)\delta(\boldsymbol{z}-\boldsymbol{z}_i) \qquad (\textit{empirical distribution})$$

$$D(\hat{p}||p_{\boldsymbol{\theta}}) = \int \mathrm{d}t\mathrm{d}\boldsymbol{z}\; \hat{p}(t,\boldsymbol{z}) \log\Big[\frac{\hat{p}(t|\boldsymbol{z})}{p(t|\boldsymbol{z},\boldsymbol{\theta})}\Big] \quad (KL-\textit{distance})$$

- so ML regression pushes $p(t|\boldsymbol{z},\boldsymbol{\theta})$
  as close as possible towards $\hat{p}(t|\boldsymbol{z})$

  true pars: $\boldsymbol{\theta}^\star$

  - fixed p, $N\to\infty$: $\quad \hat{p}(t,\boldsymbol{z}) = p(t,\boldsymbol{z}|\boldsymbol{\theta}^\star)$, $\quad$ so $\boldsymbol{\theta}_{\mathrm{ML}} = \boldsymbol{\theta}^\star$ ✓
  - $p=\mathcal{O}(N)$, $N\to\infty$: $\quad \hat{p}(t,\boldsymbol{z}) \neq p(t,\boldsymbol{z}|\boldsymbol{\theta}^\star)$, $\quad$ so $\boldsymbol{\theta}_{\mathrm{ML}} \neq \boldsymbol{\theta}^\star$ ✗

The basic ideas

Step1 – identify quantity to calculate

- $\hat{p}_{\theta^\star}$: empirical distr of $(t, \boldsymbol{z})$,
  for data generated with $\boldsymbol{\theta}^\star$

  ML regression: *minimize* $D(\hat{p}_{\theta^\star} || p_{\boldsymbol{\theta}})$

  optimal stopping point: $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$

  $D(\hat{p}_{\theta^\star} || p_{\boldsymbol{\theta}}) = D(\hat{p}_{\theta^\star} || p_{\boldsymbol{\theta}^\star})$ ← *zero iff* $p \ll N$

  define:

$$E(\boldsymbol{\theta}^\star, \mathcal{D}) = \min_{\boldsymbol{\theta}} D(\hat{p}_{\theta^\star} || p_{\boldsymbol{\theta}}) - D(\hat{p}_{\theta^\star} || p_{\boldsymbol{\theta}^\star})$$

$$E(\boldsymbol{\theta}^\star, \mathcal{D}) > 0: \text{ underfitting}$$

$$E(\boldsymbol{\theta}^\star, \mathcal{D}) < 0: \text{ overfitting}$$

- *Typical behaviour*

$$E(\boldsymbol{\theta}^\star) = \Big\langle E(\boldsymbol{\theta}^\star, \mathcal{D}) \Big\rangle_{\mathcal{D}}$$

$$= \Big\langle \min_{\boldsymbol{\theta}} \Big\{ \frac{1}{N} \sum_i \log \Big[\frac{p(t_i | \boldsymbol{z}_i, \boldsymbol{\theta}^\star)}{p(t_i | \boldsymbol{z}_i, \boldsymbol{\theta})} \Big] \Big\} \Big\rangle_{\mathcal{D}}$$

The basic ideas

Step1 – identify quantity to calculate

- $\hat{p}_{\boldsymbol{\theta}^\star}$: empirical distr of $(t, \boldsymbol{z})$,
  for data generated with $\boldsymbol{\theta}^\star$

$$\begin{aligned} &\textit{ML regression}: && \textit{minimize}\ \ D(\hat{p}_{\boldsymbol{\theta}^\star} || p_{\boldsymbol{\theta}}) \\ &\textit{optimal stopping point}: && \boldsymbol{\theta} = \boldsymbol{\theta}^\star \\ & && D(\hat{p}_{\boldsymbol{\theta}^\star} || p_{\boldsymbol{\theta}}) = D(\hat{p}_{\boldsymbol{\theta}^\star} || p_{\boldsymbol{\theta}^\star}) \quad \leftarrow \textit{zero iff } p \ll N \end{aligned}$$

  define:

$$E(\boldsymbol{\theta}^\star, \mathcal{D}) \;=\; \min_{\boldsymbol{\theta}} D(\hat{p}_{\boldsymbol{\theta}^\star} || p_{\boldsymbol{\theta}}) - D(\hat{p}_{\boldsymbol{\theta}^\star} || p_{\boldsymbol{\theta}^\star})$$

$$\begin{aligned} E(\boldsymbol{\theta}^\star, \mathcal{D}) > 0&: \ \text{underfitting} \\ E(\boldsymbol{\theta}^\star, \mathcal{D}) < 0&: \ \text{overfitting} \end{aligned}$$

- *Typical behaviour*

$$\begin{aligned} E(\boldsymbol{\theta}^\star) \;&=\; \Big\langle E(\boldsymbol{\theta}^\star, \mathcal{D}) \Big\rangle_{\mathcal{D}} \\ &=\; \Big\langle \min_{\boldsymbol{\theta}} \Big\{ \frac{1}{N} \sum_i \log \Big[\frac{p(t_i | \boldsymbol{z}_i, \boldsymbol{\theta}^\star)}{p(t_i | \boldsymbol{z}_i, \boldsymbol{\theta})} \Big] \Big\} \Big\rangle_{\mathcal{D}} \end{aligned}$$

The basic ideas

Step1 – identify quantity to calculate

- $\hat{p}_{\boldsymbol{\theta}^\star}$: empirical distr of $(t, \boldsymbol{z})$,
  for data generated with $\boldsymbol{\theta}^\star$

$$\begin{aligned} &\textit{ML regression}: && \textit{minimize}\ \ D(\hat{p}_{\boldsymbol{\theta}^\star}||p_{\boldsymbol{\theta}}) \\ &\textit{optimal stopping point}: && \boldsymbol{\theta} = \boldsymbol{\theta}^\star \\ & && D(\hat{p}_{\boldsymbol{\theta}^\star}||p_{\boldsymbol{\theta}}) = D(\hat{p}_{\boldsymbol{\theta}^\star}||p_{\boldsymbol{\theta}^\star}) \quad \leftarrow \textit{zero iff } p \ll N \end{aligned}$$

  define:

$$E(\boldsymbol{\theta}^\star, \mathcal{D}) \;=\; \min_{\boldsymbol{\theta}} D(\hat{p}_{\boldsymbol{\theta}^\star}||p_{\boldsymbol{\theta}}) - D(\hat{p}_{\boldsymbol{\theta}^\star}||p_{\boldsymbol{\theta}^\star})$$

$$E(\boldsymbol{\theta}^\star, \mathcal{D}) > 0: \ \text{underfitting}$$
$$E(\boldsymbol{\theta}^\star, \mathcal{D}) < 0: \ \text{overfitting}$$

- *Typical behaviour*

$$\begin{aligned} E(\boldsymbol{\theta}^\star) \;&=\; \Big\langle E(\boldsymbol{\theta}^\star, \mathcal{D}) \Big\rangle_{\mathcal{D}} \\ &=\; \Big\langle \min_{\boldsymbol{\theta}} \Big\{ \frac{1}{N} \sum_i \log \Big[\frac{p(t_i|\boldsymbol{z}_i, \boldsymbol{\theta}^\star)}{p(t_i|\boldsymbol{z}_i, \boldsymbol{\theta})} \Big] \Big\} \Big\rangle_{\mathcal{D}} \end{aligned}$$

Step 2 – remove minimisation over $\boldsymbol{\theta}$

$$E(\boldsymbol{\theta}^\star) = \Big\langle \min_{\boldsymbol{\theta}} \Big\{ \frac{1}{N} \sum_i \log \Big[\frac{p(t_i|\boldsymbol{z}_i, \boldsymbol{\theta}^\star)}{p(t_i|\boldsymbol{z}_i, \boldsymbol{\theta})} \Big] \Big\} \Big\rangle_{\mathcal{D}}$$

- *Laplace identity*

$$\lim_{\gamma\to\infty} \frac{\partial}{\partial\gamma} \log \int \mathrm{d}x\ \mathrm{e}^{\gamma f(x)} \;=\; \lim_{\gamma\to\infty} \frac{\int \mathrm{d}x\ \mathrm{e}^{\gamma f(x)} f(x)}{\int \mathrm{d}x\ \mathrm{e}^{\gamma f(x)}} \;=\; \max_x f(x)$$

use in reverse:

$$E(\boldsymbol{\theta}^\star) \;=\; -\lim_{\gamma\to\infty} \frac{1}{N} \frac{\partial}{\partial\gamma} \Big\langle \log \int \mathrm{d}\boldsymbol{\theta} \prod_{i=1}^{N} \Big[\frac{p(t_i|\boldsymbol{z}_i, \boldsymbol{\theta})}{p(t_i|\boldsymbol{z}_i, \boldsymbol{\theta}^\star)} \Big]^\gamma \Big\rangle_{\mathcal{D}}$$

interpretation:
stochastic minimisation, with noise $\sim 1/\gamma$

Step 2 – remove minimisation over $\boldsymbol{\theta}$

$$E(\boldsymbol{\theta}^\star) = \Big\langle \min_{\boldsymbol{\theta}} \Big\{ \frac{1}{N} \sum_i \log \Big[\frac{p(t_i|\boldsymbol{z}_i, \boldsymbol{\theta}^\star)}{p(t_i|\boldsymbol{z}_i, \boldsymbol{\theta})} \Big] \Big\} \Big\rangle_{\mathcal{D}}$$

- *Laplace identity*

$$\lim_{\gamma\to\infty} \frac{\partial}{\partial\gamma} \log \int \mathrm{d}x\, \mathrm{e}^{\gamma f(x)} \;=\; \lim_{\gamma\to\infty} \frac{\int \mathrm{d}x\, \mathrm{e}^{\gamma f(x)} f(x)}{\int \mathrm{d}x\, \mathrm{e}^{\gamma f(x)}} \;=\; \max_x f(x)$$

  use in reverse:

$$E(\boldsymbol{\theta}^\star) \;=\; -\lim_{\gamma\to\infty} \frac{1}{N} \frac{\partial}{\partial\gamma} \Big\langle \log \int \mathrm{d}\boldsymbol{\theta} \prod_{i=1}^{N} \Big[\frac{p(t_i|\boldsymbol{z}_i, \boldsymbol{\theta})}{p(t_i|\boldsymbol{z}_i, \boldsymbol{\theta}^\star)} \Big]^\gamma \Big\rangle_{\mathcal{D}}$$

interpretation:
stochastic minimisation, with noise $\sim 1/\gamma$

The replica method

aim: make hard analytical calculations easy ...
here: compute the average over $\mathcal{D}$

$$E(\boldsymbol{\theta}^\star) \quad = \quad -\lim_{\gamma\to\infty}\frac{1}{N}\frac{\partial}{\partial\gamma}\Big\langle \log\int \mathrm{d}\boldsymbol{\theta}\,\prod_{i=1}^{N}\Big[\frac{p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta})}{p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta}^\star)}\Big]^\gamma\Big\rangle_{\mathcal{D}}$$

- replica method
  – evaluate for *integer* n,
  – analytical continuation to *non-integer* n

$$\langle \log Z\rangle \;=\; \lim_{n\to 0}\frac{1}{n}\log\langle Z^n\rangle$$

- application

$$E(\theta^\star) \quad = \quad -\lim_{\gamma\to\infty}\frac{1}{N}\frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{n}\log\Big\langle\Big[\int \mathrm{d}\theta\,\prod_{i=1}^{N}\Big[\frac{p(t_i|\boldsymbol{z}_i,\theta)}{p(t_i|\boldsymbol{z}_i,\theta^\star)}\Big]^\gamma\Big]^n\Big\rangle_{\mathcal{D}}$$

$$= \; -\lim_{\gamma\to\infty}\lim_{n\to 0}\frac{1}{Nn}\frac{\partial}{\partial\gamma}\log\int \mathrm{d}\theta^1\ldots\mathrm{d}\theta^n\Big[\int \mathrm{d}\boldsymbol{z}\mathrm{d}t\; p(\boldsymbol{z})p(t|\boldsymbol{z},\theta^\star)\prod_{\alpha=1}^{n}\Big[\frac{p(t|\boldsymbol{z},\theta^\alpha)}{p(t|\boldsymbol{z},\theta^\star)}\Big]^\gamma\Big]^N$$

The replica method

aim: make hard analytical calculations easy ...
here: compute the average over $\mathcal{D}$

$$E(\boldsymbol{\theta}^\star) \;=\; -\lim_{\gamma\to\infty}\frac{1}{N}\frac{\partial}{\partial\gamma}\Big\langle \log\int \mathrm{d}\boldsymbol{\theta}\,\prod_{i=1}^{N}\Big[\frac{p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta})}{p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta}^\star)}\Big]^\gamma\Big\rangle_{\mathcal{D}}$$

- replica method
  – evaluate for *integer n*,
  – analytical continuation to *non-integer n*

$$\langle \log Z\rangle \;=\; \lim_{n\to 0}\frac{1}{n}\log\langle Z^n\rangle$$

- application

$$E(\theta^\star) \;=\; -\lim_{\gamma\to\infty}\frac{1}{N}\frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{n}\log\Big\langle\Big[\int \mathrm{d}\theta\,\prod_{i=1}^{N}\Big[\frac{p(t_i|\boldsymbol{z}_i,\theta)}{p(t_i|\boldsymbol{z}_i,\theta^\star)}\Big]^\gamma\Big]^n\Big\rangle_{\mathcal{D}}$$

$$= \;-\lim_{\gamma\to\infty}\lim_{n\to 0}\frac{1}{Nn}\frac{\partial}{\partial\gamma}\log\int \mathrm{d}\theta^1\ldots\mathrm{d}\theta^n\Big[\int \mathrm{d}\boldsymbol{z}\mathrm{d}t\; p(\boldsymbol{z})p(t|\boldsymbol{z},\theta^\star)\prod_{\alpha=1}^{n}\Big[\frac{p(t|\boldsymbol{z},\theta^\alpha)}{p(t|\boldsymbol{z},\theta^\star)}\Big]^\gamma\Big]^N$$

The replica method

aim: make hard analytical calculations easy ...
here: compute the average over $\mathcal{D}$

$$E(\boldsymbol{\theta}^\star) \;=\; -\lim_{\gamma\to\infty}\frac{1}{N}\frac{\partial}{\partial\gamma}\Big\langle \log\int \mathrm{d}\boldsymbol{\theta}\,\prod_{i=1}^{N}\Big[\frac{p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta})}{p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta}^\star)}\Big]^\gamma\Big\rangle_{\mathcal{D}}$$

- replica method
  – evaluate for *integer n*,
  – analytical continuation to *non-integer n*

$$\langle \log Z\rangle \;=\; \lim_{n\to 0}\frac{1}{n}\log\langle Z^n\rangle$$

- application

$$E(\boldsymbol{\theta}^\star) \;=\; -\lim_{\gamma\to\infty}\frac{1}{N}\frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{n}\log\Big\langle\Big[\int \mathrm{d}\boldsymbol{\theta}\,\prod_{i=1}^{N}\Big[\frac{p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta})}{p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta}^\star)}\Big]^\gamma\Big]^n\Big\rangle_{\mathcal{D}}$$

$$= \;-\lim_{\gamma\to\infty}\lim_{n\to 0}\frac{1}{Nn}\frac{\partial}{\partial\gamma}\log\int \mathrm{d}\boldsymbol{\theta}^1\ldots\mathrm{d}\boldsymbol{\theta}^n\Big[\int \mathrm{d}\boldsymbol{z}\mathrm{d}t\; p(\boldsymbol{z})p(t|\boldsymbol{z},\boldsymbol{\theta}^\star)\prod_{\alpha=1}^{n}\Big[\frac{p(t|\boldsymbol{z},\boldsymbol{\theta}^\alpha)}{p(t|\boldsymbol{z},\boldsymbol{\theta}^\star)}\Big]^\gamma\Big]^N$$

Application to Cox regression

- explicit formula for $E(S, \lambda^\star, \zeta)$

  $\zeta = p/N$, $S = |\beta^\star|$

  $\beta^\star, \lambda^\star(t)$: true associations and base hazard rate

- requires solving $u, v, w, \lambda(t)$ from

$$\begin{aligned}
\zeta v^2 &= \int \mathrm{D}z\mathrm{D}y \int \mathrm{d}t\; p(t|Sy,\lambda^\star)\Big[u^2 - W\big(u^2 \mathrm{e}^{u^2+wy+vz}\Lambda(t)\big)\Big]^2 \\
\zeta &= \int \mathrm{D}z\mathrm{D}y \int \mathrm{d}t\; p(t|Sy,\lambda^\star)\frac{W\big(u^2 \mathrm{e}^{u^2+wy+vz}\Lambda(t)\big)}{1+W\big(u^2 \mathrm{e}^{u^2+wy+vz}\Lambda(t)\big)} \\
0 &= \int \mathrm{D}z\mathrm{D}y\; y \int \mathrm{d}t\; p(t|Sy,\lambda^\star) W\big(u^2 \mathrm{e}^{u^2+wy+vz}\Lambda(t)\big) \\
\frac{p(t)}{\lambda(t)} &= \int \mathrm{D}z\mathrm{D}y \int_t^\infty \mathrm{d}t'\; p(t'|Sy,\lambda^\star)\frac{W\big(u^2 \mathrm{e}^{u^2+wy+vz}\Lambda(t')\big)}{u^2\Lambda(t')}
\end{aligned}$$

$$\mathrm{D}z = (2\pi)^{-1/2}\mathrm{e}^{-\frac{1}{2}z^2}\mathrm{d}z$$

W: Lambert function

$$p(t|\xi,\lambda) = \lambda(t)\mathrm{e}^{\xi - \exp(\xi)\Lambda(t)}$$

Application to Cox regression

- explicit formula for $E(S, \lambda^\star, \zeta)$

  $\zeta = p/N$, $S = |\beta^\star|$

  $\beta^\star, \lambda^\star(t)$: true associations and base hazard rate

- requires solving $u, v, w, \lambda(t)$ from

$$\begin{aligned}
\zeta v^2 &= \int \mathrm{D}z\mathrm{D}y \int \mathrm{d}t\; p(t|Sy,\lambda^\star)\Big[u^2 - W\big(u^2 \mathrm{e}^{u^2+wy+vz}\Lambda(t)\big)\Big]^2 \\
\zeta &= \int \mathrm{D}z\mathrm{D}y \int \mathrm{d}t\; p(t|Sy,\lambda^\star)\frac{W\big(u^2 \mathrm{e}^{u^2+wy+vz}\Lambda(t)\big)}{1+W\big(u^2 \mathrm{e}^{u^2+wy+vz}\Lambda(t)\big)} \\
0 &= \int \mathrm{D}z\mathrm{D}y\; y \int \mathrm{d}t\; p(t|Sy,\lambda^\star) W\big(u^2 \mathrm{e}^{u^2+wy+vz}\Lambda(t)\big) \\
\frac{p(t)}{\lambda(t)} &= \int \mathrm{D}z\mathrm{D}y \int_t^\infty \mathrm{d}t'\; p(t'|Sy,\lambda^\star)\frac{W\big(u^2 \mathrm{e}^{u^2+wy+vz}\Lambda(t')\big)}{u^2\Lambda(t')}
\end{aligned}$$

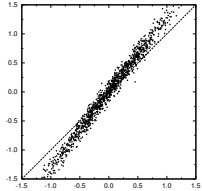
$$\mathrm{D}z = (2\pi)^{-1/2}\mathrm{e}^{-\frac{1}{2}z^2}\mathrm{d}z$$

W: Lambert function

$$p(t|\xi,\lambda) = \lambda(t)\mathrm{e}^{\xi - \exp(\xi)\Lambda(t)}$$

- interpretation:

$$\textit{slope}: \quad \kappa = w/S$$
$$\textit{width}: \quad \sigma = v/\sqrt{p}$$

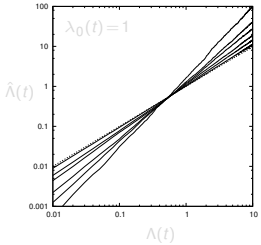


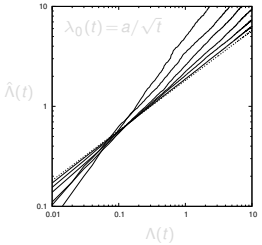
all we need for overfitting correction!

- challenge: eqn for $\Lambda(t)$

$$t \gg 1: \qquad \log \Lambda(t) = \rho \log \Lambda^\star(t) + (1-\rho) \log\log \Lambda^\star(t) + \ldots$$

$$\rho = \frac{w}{2S}\left(1+\sqrt{1+4u^2/w^2}\right)$$



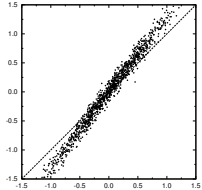


- variational approx:
  $\Lambda(t) = k[\Lambda^\star(t)]^\rho$

  $\Lambda^\star(t)$ drops out of equations!

- interpretation:

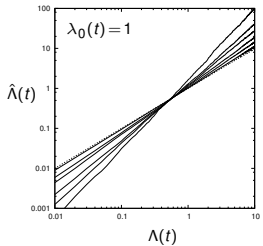
$$\textit{slope}: \quad \kappa = w/S$$
$$\textit{width}: \quad \sigma = v/\sqrt{p}$$

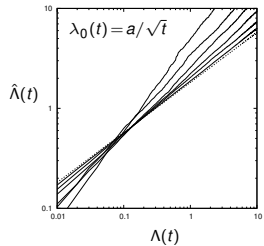


all we need for overfitting correction!

- challenge: eqn for $\Lambda(t)$

$$t \gg 1: \qquad \log \Lambda(t) = \rho \log \Lambda^\star(t) + (1-\rho) \log\log \Lambda^\star(t) + \ldots$$
$$\rho = \frac{w}{2S}\left(1+\sqrt{1+4u^2/w^2}\right)$$





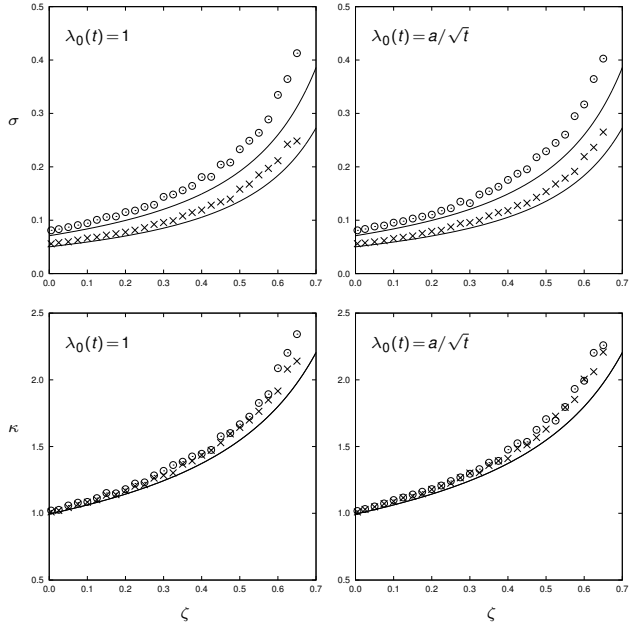
- variational approx:
  $\Lambda(t) = k[\Lambda^\star(t)]^\rho$

  $\Lambda^\star(t)$ drops out of equations!

width σ and slope κ
of data clouds

lines: variational theory
for $S = 0.5$ and $\langle t \rangle = 1$

simulations:
o: $N = 200$
x: $N = 400$



Application to regularized Cox regression

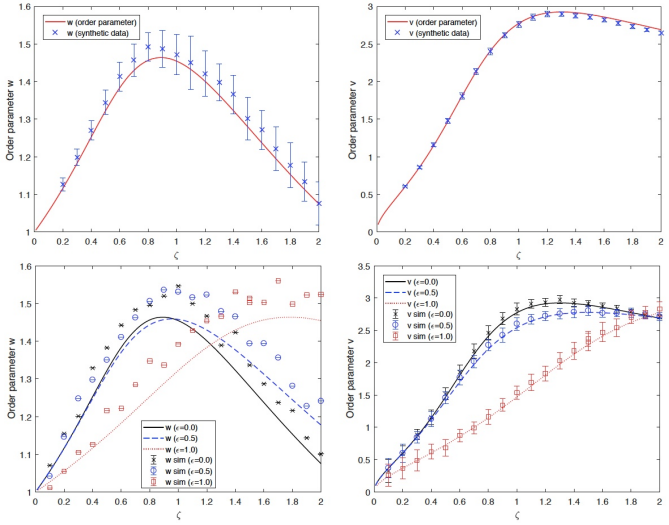
$N, p \to \infty,\ \zeta = p/N,$
L2-prior: $p(\beta) \propto \exp(-\eta p \beta^2)$

main changes:

- dependence of theory on eigenvalue spectrum $\varrho(a)$ of covariate correlation matrix
- two extra order parameters, closed equations for $u, v, w, f, g, \lambda(t)$
- no longer a phase transition at $\zeta = 1$, but inference well-defined for any $\zeta > 0$
- closed eqn for *optimal* hyperparameter η, defined by demanding absent bias, i.e. slope $\kappa = 1$

uncorrelated covariates

pairwise correlated covariates



$\eta = 0.025$, top row $p = 2000$, bottom row $Np = 400{,}000$

overfitting correction pars: slope $\kappa = w/\tilde{S}$, width $\sigma = v$

final overfitting correction protocol

1. estimate covariate correlation matrix $\boldsymbol{A}$,
   compute its eigenvalue spectrum $\varrho(a)$
2. carry out regularized Cox regression,
   with prior $p(\boldsymbol{\beta}) \propto \exp(-\eta p \boldsymbol{\beta}^2)$ (small $\eta > 0$)
   result: $\hat{\boldsymbol{\beta}}$ and $\hat{\Lambda}(t)$ (Breslow estimator)
3. calculate $\hat{\boldsymbol{\beta}} \cdot \boldsymbol{A}\hat{\boldsymbol{\beta}}$
4. solve coupled nonlinear eqns for (u, v, w, f, g, k, ρ),
   alongside $v^2 + w^2 = \hat{\boldsymbol{\beta}} \cdot \boldsymbol{A}\hat{\boldsymbol{\beta}}$
   (replaces unknown variable S)
5. calculate slope κ and noise amplitude σ
6. compute corrected estimators:

$$\hat{\boldsymbol{\beta}} = \kappa^{-1}\hat{\boldsymbol{\beta}}, \qquad \hat{\Lambda}(t) = [\hat{\Lambda}(t)/k]^{1/\rho}$$

7. use σ to correct p-values

Other extensions

- survival analysis with censoring
  - – end-of-trial censoring: minimal change to the theory
  - – censoring by competing risks: more complex but doable
- generalized linear models:

$$p(y|\boldsymbol{z}) = p(y|\beta^1 \cdot \boldsymbol{z}, \ldots, \beta^K \cdot \boldsymbol{z}; \omega)$$

e.g.

Logistic regression

$$y \in \{-1, 1\} \qquad p(y|\boldsymbol{z}) = \frac{e^{y(\beta \cdot \boldsymbol{z} + \beta_0)}}{2\cosh(\beta \cdot \boldsymbol{z} + \beta_0)}$$

Ordinal class regression

$$y \in \{1, \ldots, C\} \qquad p(y|\boldsymbol{z}) = e^{-\exp(\beta \cdot \boldsymbol{z}) \sum_{y' > y} \lambda_{y'}} - e^{-\exp(\beta \cdot \boldsymbol{z}) \sum_{y' \geq y} \lambda_{y'}}$$

Latent class survival analysis

$$t > 0 \qquad p(t|\boldsymbol{z}) = \sum_{\ell=1}^{L} w_\ell \left[\lambda_\ell(t) e^{\beta^\ell \cdot \boldsymbol{z} - \exp(\beta^\ell \cdot \boldsymbol{z}) \int_0^t dt' \lambda_\ell(t')} \right]$$

Other extensions

- survival analysis with censoring
  - – end-of-trial censoring: minimal change to the theory
  - – censoring by competing risks: more complex but doable
- generalized linear models:

$$p(y|\boldsymbol{z}) = p(y|\boldsymbol{\beta}^1 \cdot \boldsymbol{z}, \ldots, \boldsymbol{\beta}^K \cdot \boldsymbol{z}; \boldsymbol{\omega})$$

e.g.

Logistic regression

$$y \in \{-1, 1\} \qquad p(y|\boldsymbol{z}) = \frac{\mathrm{e}^{y(\boldsymbol{\beta}\cdot\boldsymbol{z}+\beta_0)}}{2\cosh(\boldsymbol{\beta}\cdot\boldsymbol{z}+\beta_0)}$$

Ordinal class regression

$$y \in \{1, \ldots, C\} \qquad p(y|\boldsymbol{z}) = \mathrm{e}^{-\exp(\boldsymbol{\beta}\cdot\boldsymbol{z})\sum_{y'>y}\lambda_{y'}} - \mathrm{e}^{-\exp(\boldsymbol{\beta}\cdot\boldsymbol{z})\sum_{y'\geq y}\lambda_{y'}}$$

Latent class survival analysis

$$t > 0 \qquad p(t|\boldsymbol{z}) = \sum_{\ell=1}^{L} w_\ell \left[\lambda_\ell(t)\mathrm{e}^{\boldsymbol{\beta}^\ell\cdot\boldsymbol{z}-\exp(\boldsymbol{\beta}^\ell\cdot\boldsymbol{z})\int_0^t \mathrm{d}t'\,\lambda_\ell(t')}\right]$$

Summary

- *Overfitting in Cox regression*

  (i) bias in regression parameters: $\hat{\boldsymbol{\beta}} \approx \kappa \boldsymbol{\beta}^{\star}$, $\kappa > 1$
  (ii) bias in base hazard rates
  (iii) extra noise, not captured by p-values

- *Analytical approach based on the replica method*

  nontrivial closed equations for $\{u, v, w, \lambda(t)\}$
  variational approximation for $\lambda(t)$

  predictions for slope κ and noise σ
  phase transition at $p/N = 1$

- *L2-regularized Cox regression*

  theory involves spectrum of covariate correlation matrix
  p-dependent hyperparameter
  phase transition removed

reliable basis for overfitting corrections,
easily extended to arbitrary generalized linear models

https://nms.kcl.ac.uk/ton.coolen

Thank you!

Summary

- *Overfitting in Cox regression*

  (i) bias in regression parameters: $\hat{\boldsymbol{\beta}} \approx \kappa \boldsymbol{\beta}^{\star}$, $\kappa > 1$
  (ii) bias in base hazard rates
  (iii) extra noise, not captured by p-values

- *Analytical approach based on the replica method*

  nontrivial closed equations for $\{u, v, w, \lambda(t)\}$
  variational approximation for $\lambda(t)$

  predictions for slope κ and noise σ
  phase transition at $p/N = 1$

- *L2-regularized Cox regression*

  theory involves spectrum of covariate correlation matrix
  p-dependent hyperparameter
  phase transition removed

reliable basis for overfitting corrections,
easily extended to arbitrary generalized linear models

https://nms.kcl.ac.uk/ton.coolen

Thank you!

Summary

- *Overfitting in Cox regression*

  (i) bias in regression parameters: $\hat{\boldsymbol{\beta}} \approx \kappa \boldsymbol{\beta}^{\star}, \ \kappa > 1$
  (ii) bias in base hazard rates
  (iii) extra noise, not captured by p-values

- *Analytical approach based on the replica method*

  nontrivial closed equations for $\{u, v, w, \lambda(t)\}$
  variational approximation for $\lambda(t)$

  predictions for slope κ and noise σ
  phase transition at $p/N = 1$

- *L2-regularized Cox regression*

  theory involves spectrum of covariate correlation matrix
  p-dependent hyperparameter
  phase transition removed

reliable basis for overfitting corrections,
easily extended to arbitrary generalized linear models

https://nms.kcl.ac.uk/ton.coolen

Thank you!

Summary

- *Overfitting in Cox regression*

  (i) bias in regression parameters: $\hat{\boldsymbol{\beta}} \approx \kappa \boldsymbol{\beta}^{\star}, \ \kappa > 1$
  (ii) bias in base hazard rates
  (iii) extra noise, not captured by p-values

- *Analytical approach based on the replica method*

  nontrivial closed equations for $\{u, v, w, \lambda(t)\}$
  variational approximation for $\lambda(t)$

  predictions for slope κ and noise σ
  phase transition at $p/N = 1$

- *L2-regularized Cox regression*

  theory involves spectrum of covariate correlation matrix
  p-dependent hyperparameter
  phase transition removed

reliable basis for overfitting corrections,
easily extended to arbitrary generalized linear models

https://nms.kcl.ac.uk/ton.coolen

Thank you!