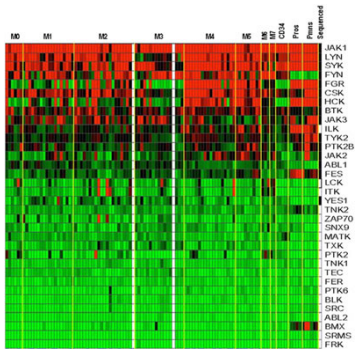
Replica analysis of overfitting in survival analysis and other generalized linear models

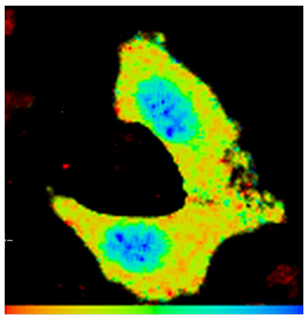
ACC Coolen

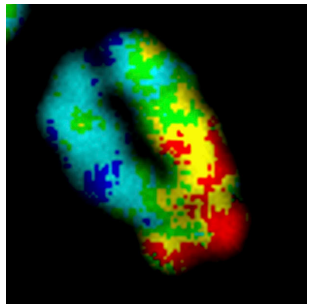
King's College London and Saddle Point Science

Sendai, Oct 27th 2019

Data of modern medicine



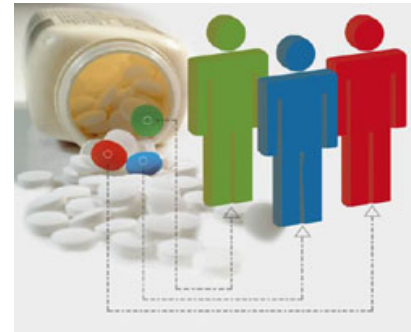




- *volume* ...
- *diversity* ...
  (clinical, genomic, biomarkers, health records, imaging, ...)
- *complexity* of pipelines ...
  (confounders, batch effects, variability between centres, ...)
- *dimensionality*

  clinical ($\sim 10^1$), metabolic ($\sim 10^2$), gene expr ($\sim 10^5$),
  imaging ($\sim 10^6$), NGS ($\sim 10^{10}$ and more)

generating ‘big data’ is not enough ...



- ‘right drug, right dose, at right time ...’

  need *predictive models* $p(y|\boldsymbol{z})$,

  $\boldsymbol{z}$: mutations, gene expr, biomarkers, images, ...
  y: treatment response

- regression:
  find parameters $\boldsymbol{\theta}$ of model $p(y|\boldsymbol{z},\boldsymbol{\theta})$ from data

  curse of dimensionality ...

  pre-genome: $N \sim 10^3$, dim $\boldsymbol{\theta} \sim 10^2$
  post-genome: $N \sim 10^4$, dim $\boldsymbol{\theta} \sim 10^{10}$ ...

- simpler question: predict *individual* risk
  (target aggressive treatments)

  cancer: outcome is often a *duration* t,
  (overall or progression-free survival)

  predictive model: $p(t|\boldsymbol{z},\boldsymbol{\theta})$

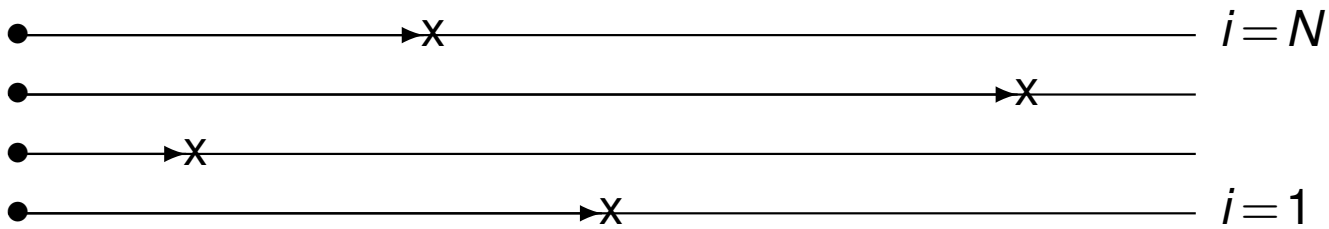
Regression for time-to-event data

- *Data* $\mathcal{D} = \{(\boldsymbol{z}_1, t_1), \ldots, (\boldsymbol{z}_N, t_N)\}$

  samples $(\boldsymbol{z}_i, t_i)$

$\boldsymbol{z}_i$: *p covariates* (*measured at baseline*)

$t_i > 0$: *failure time* (*death*, *relapse*, ...)



- *Aim*

  find and quantify patterns that relate
  covariates to event times, in order to:

  1. *identify high-risk patients*
  2. *discover disease mechanisms*
  3. *design new treatments (modifiable covariates)*

Proportional hazards regression (1972)

main tool of medical statistics

$$\textit{event time dist}: \quad p(t|\boldsymbol{z}) = h(t|\boldsymbol{z}) \mathrm{e}^{-\int_0^t \mathrm{d}t' \, h(t'|\boldsymbol{z})}$$

$$\textit{hazard rate}: \quad h(t|\boldsymbol{z}) = \lambda(t) \mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}}$$

parameters: $\boldsymbol{\beta}, \lambda(t)$

- Maximum likelihood:

$$(\hat{\boldsymbol{\beta}}, \hat{\lambda}) \quad = \quad \mathrm{argmax}_{\boldsymbol{\beta},\lambda} \Big\{ \frac{1}{N} \sum_i \log p(t_i|\boldsymbol{z}_i, \boldsymbol{\beta}, \lambda) \Big\}$$

- Maximise over $\lambda(t)$ *first*

$$\hat{\lambda}(t|\boldsymbol{\beta}) \quad = \quad \frac{\sum_j \delta(t-t_j)}{\sum_k \theta(t_k - t) \, \mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}_k}} \qquad (\textit{Breslow estimator})$$

$$\hat{\boldsymbol{\beta}} \quad = \quad \mathrm{argmax}_{\boldsymbol{\beta}} \Big\{ \sum_i \boldsymbol{\beta}\cdot\boldsymbol{z}_i - \sum_i \log \Big[\frac{\sum_j \mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}_j} \theta(t_j - t_i)}{\sum_j \theta(t_j - t_i)} \Big] \Big\}$$

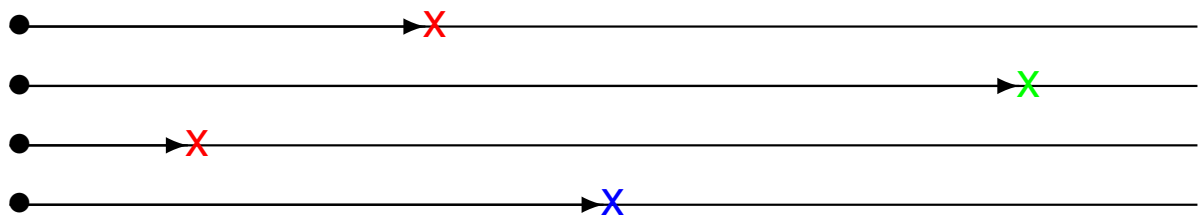
Beyond the basic model ...

- *Mathematical subtleties*

  cure is forbidden, i.e. we need $\int_0^\infty \mathrm{d}t\ \lambda(t) < \infty$,
  violated by Breslow estimator ...

  issues with event time coincidences,
  i.e. we need $t_i \neq t_j$ for all i, j

- *Multiple competing risks*

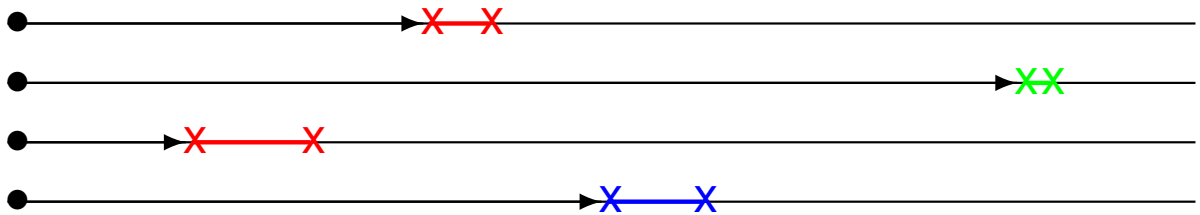


  risk labels $r_i \in \{1, \ldots, R\}$,

$$\mathcal{D} = \{(\boldsymbol{z}_1, t_1, r_1), \ldots, (\boldsymbol{z}_N, t_N, r_N)\}$$

$$p(t, r|\boldsymbol{z}, \boldsymbol{\theta}) = h_r(t|\boldsymbol{z}, \boldsymbol{\theta}) \exp\Big[-\int_0^t \mathrm{d}t' \sum_{r'=1}^{R} h_{r'}(t'|\boldsymbol{z}, \boldsymbol{\theta})\Big]$$

- *interval censoring*

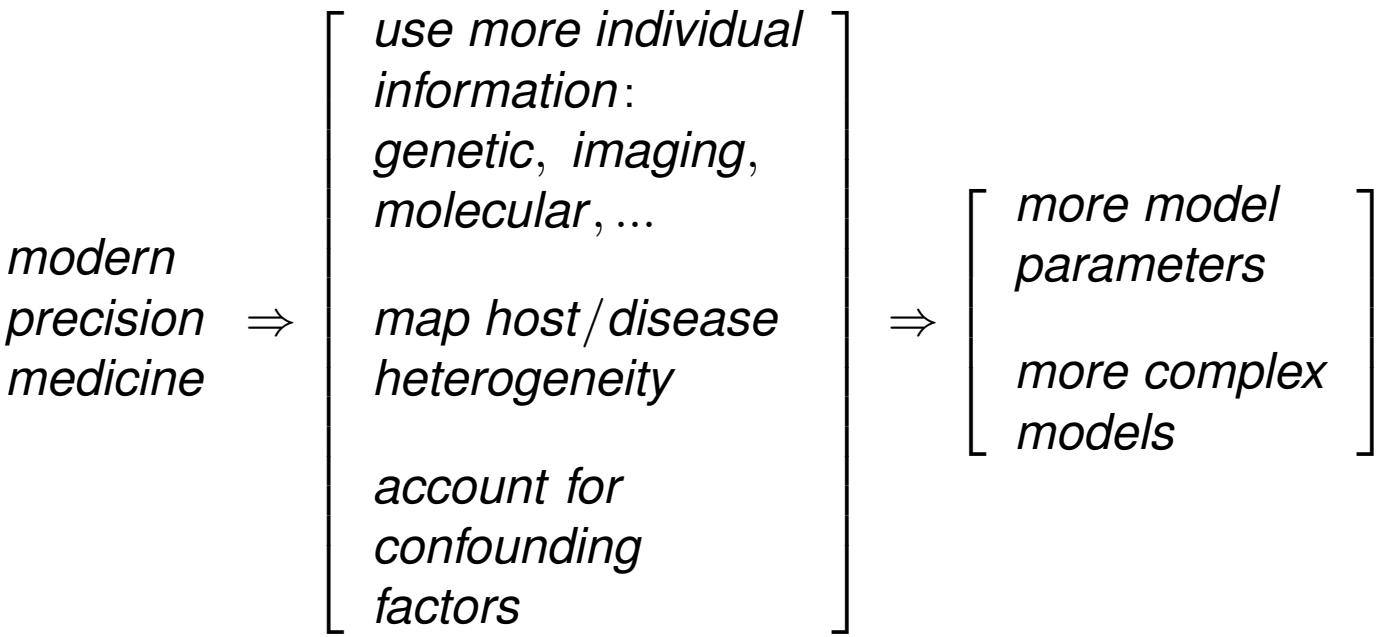


- *Include latent heterogeneity*

  (e.g. mixture models, frailty models, random effects models)

  distinct patient sub-classes, $\ell = 1 \ldots L$,

$$p(t, r|\boldsymbol{z}, \boldsymbol{\theta}) = \sum_{\ell=1}^{L} w_\ell \; h_r(t'|\boldsymbol{z}, \boldsymbol{\theta}, \ell) \exp\Big[- \int_0^t \mathrm{d}t' \sum_{r'=1}^{R} h_{r'}(t'|\boldsymbol{z}, \boldsymbol{\theta}, \ell) \Big]$$


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*

Strategies to deal with overfitting

in covariate-to-outcome analysis

- *'Back off'*

  'safe' ratio covariates/samples
  for multivariate regression

- *Eliminate redundant information*

  improve covariates/samples ratio via
  supervised nonlinear dimension reduction,
  or by using biological knowledge.
  'true' data dimension?

- *'Integrate out' overfitting effects*

  fully Bayesian analysis of parameter uncertainty,
  while keeping computation feasible

- *Model overfitting effects*

  Overfitting correction theory for
  multivariate regression

Phenomenology of overfitting

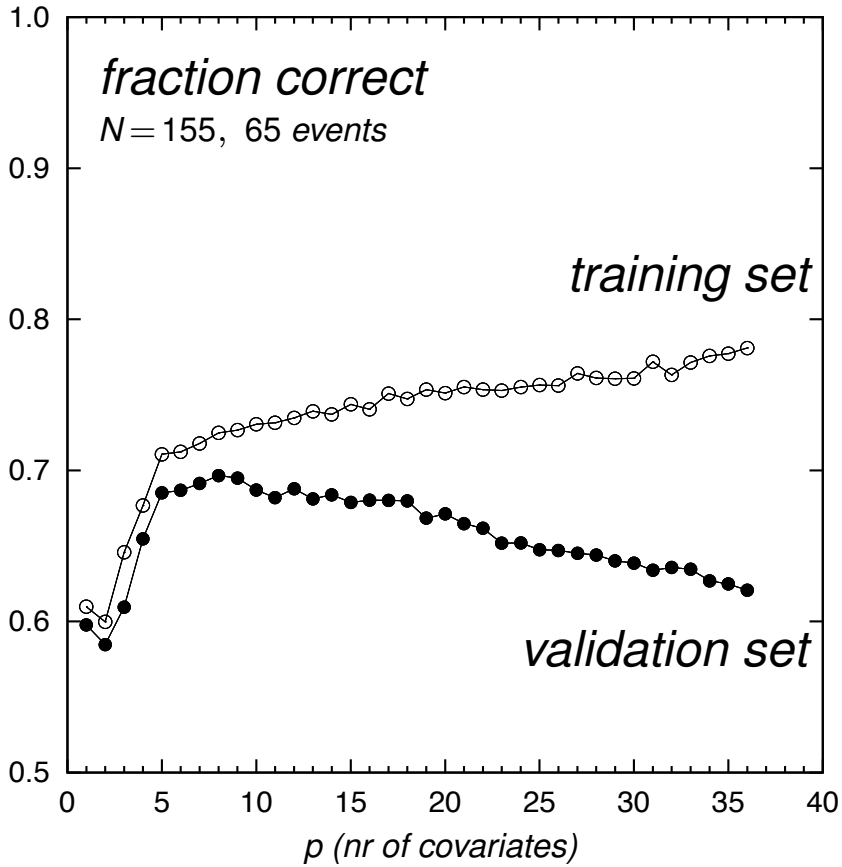
deteriorating outcome
prediction performance
on unseen data ...

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 ...

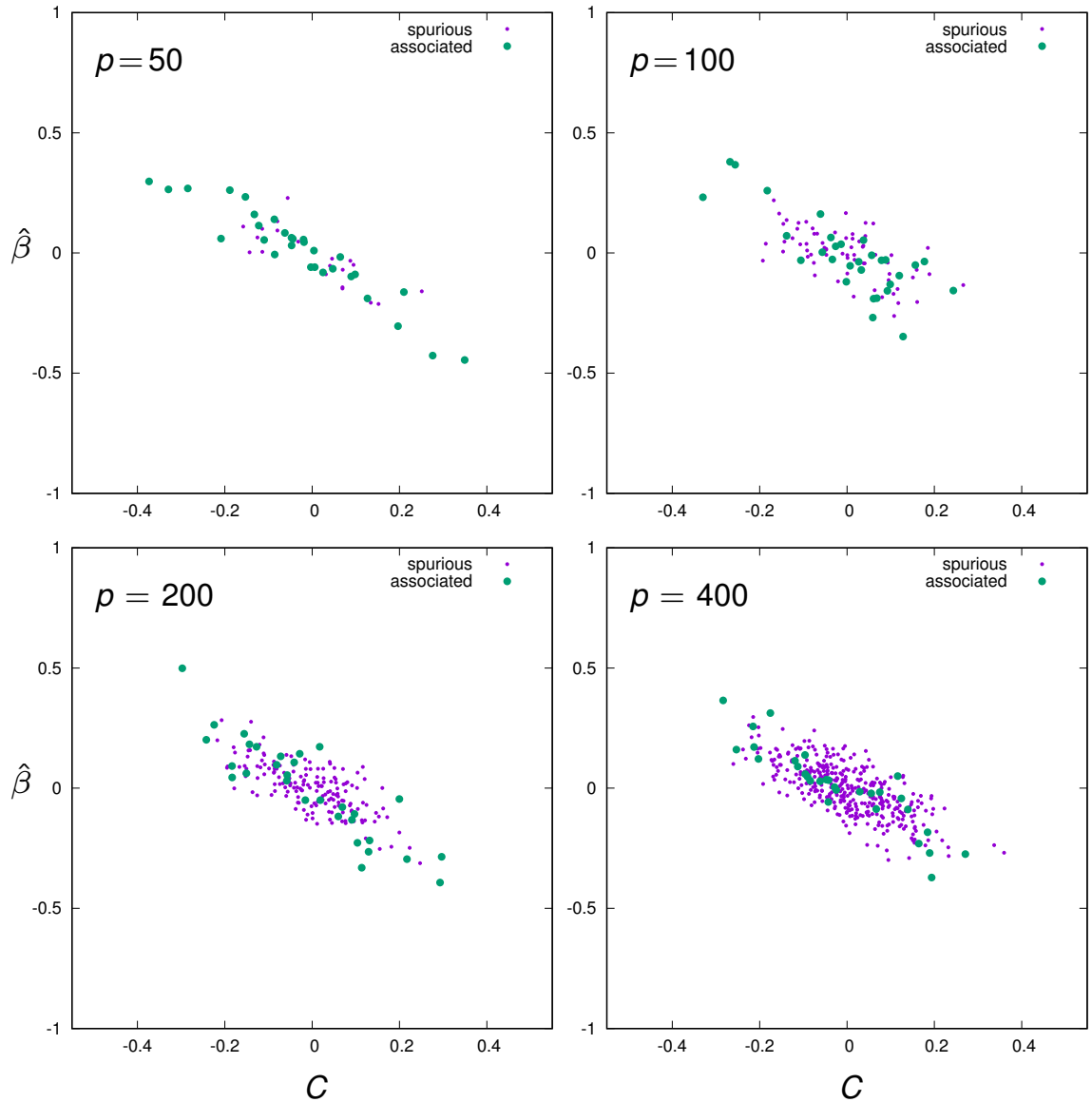
$N = 100$

p covariates:

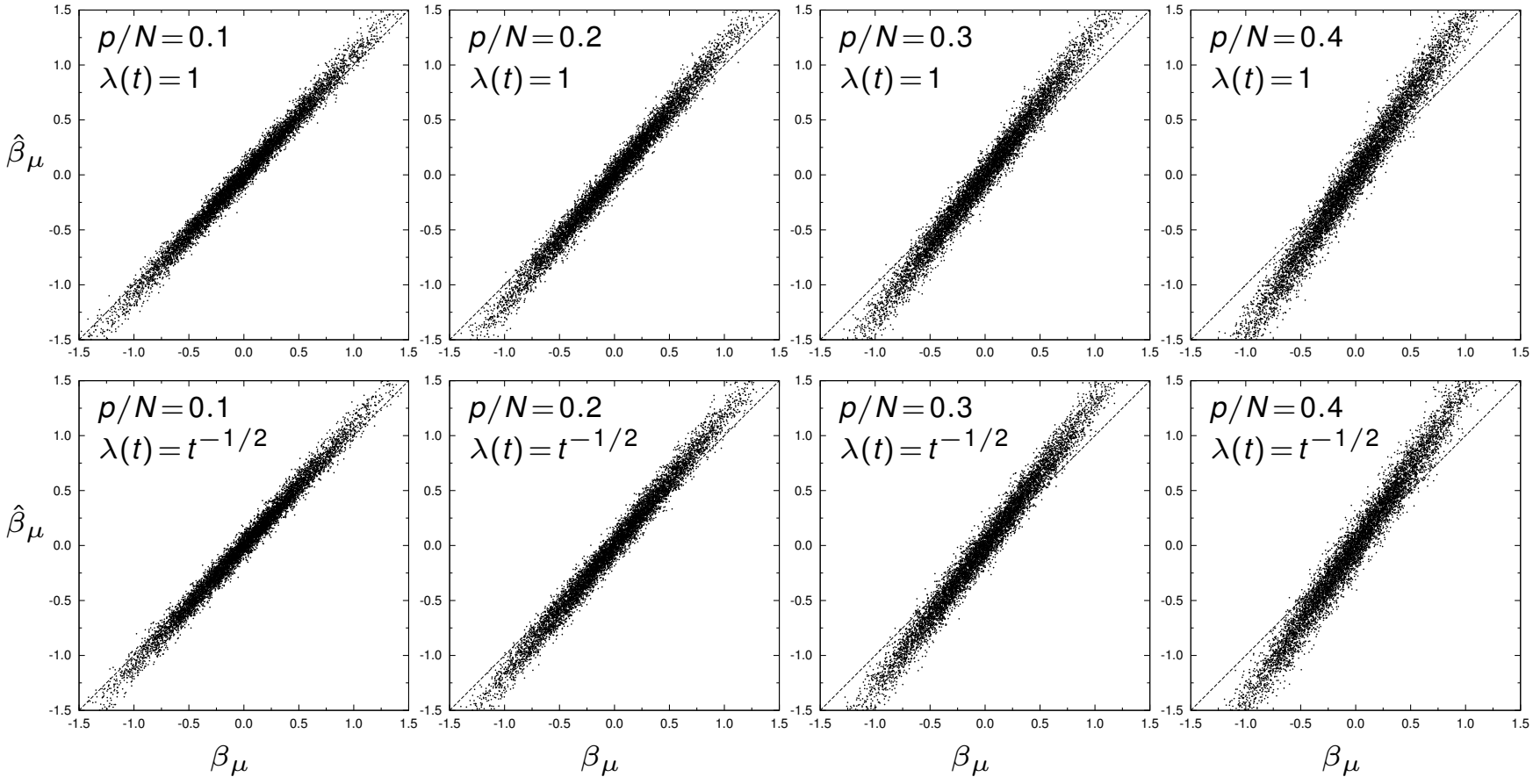
30 true associations
$p - 30$ spurious ones

C: Pearson correlation between covariates and event time,

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



bias in inferred association parameters ...



β_μ: true associations

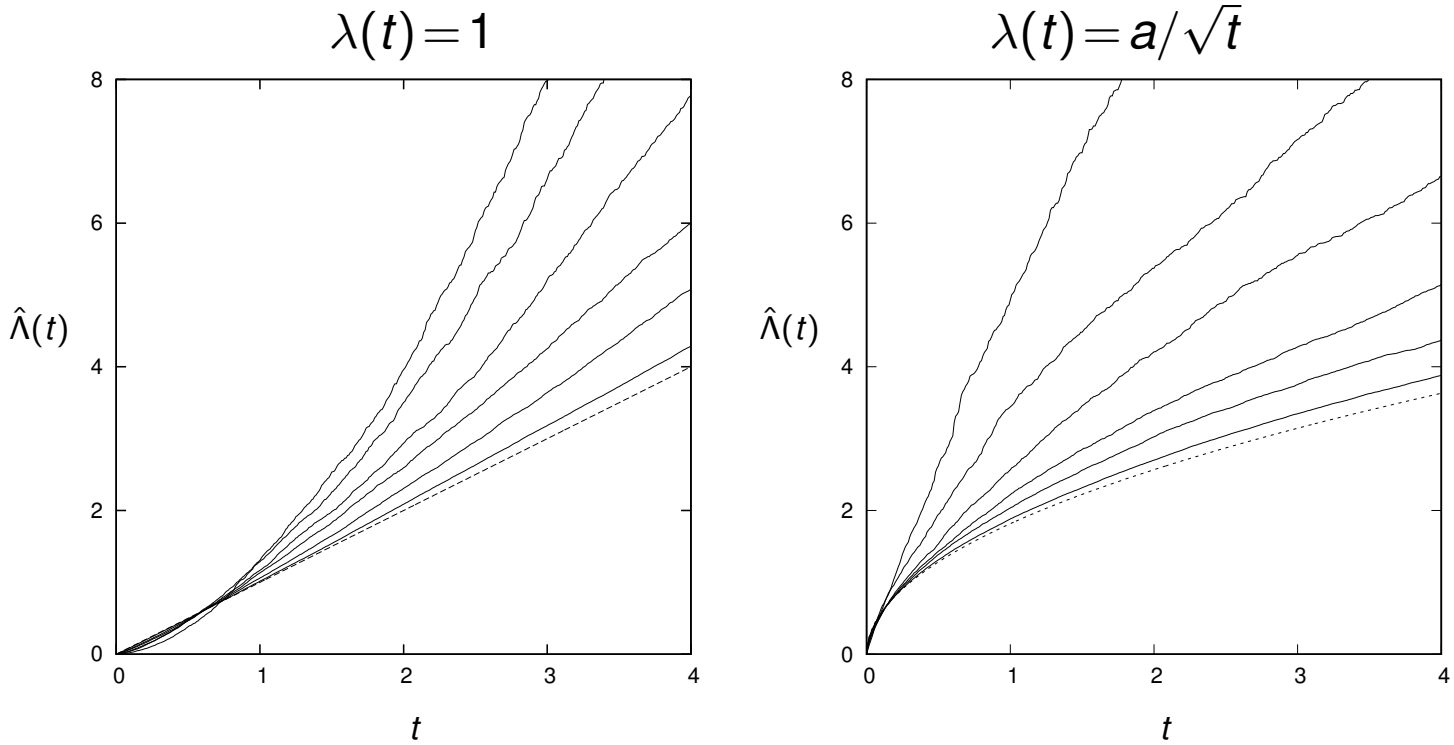
$\hat{\beta}_\mu$: multivariate regression

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

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

notation:

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

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

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

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

scaling of hyperparameters

in regularised Cox regression

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

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

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!

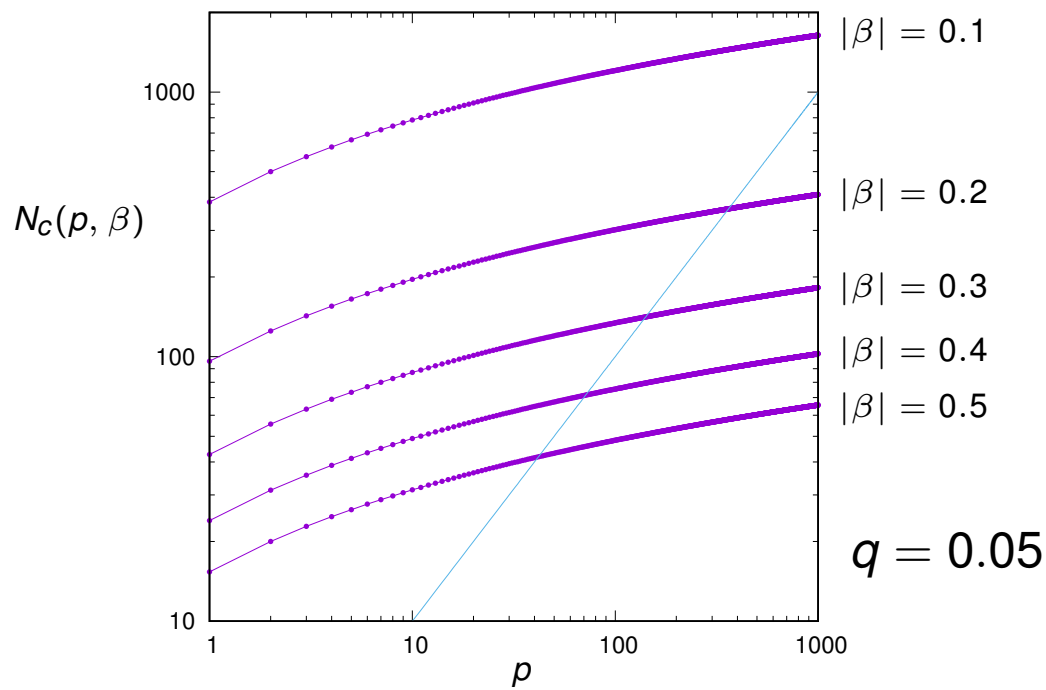
regression with $p > N$ in principle possible

provided N is large enough

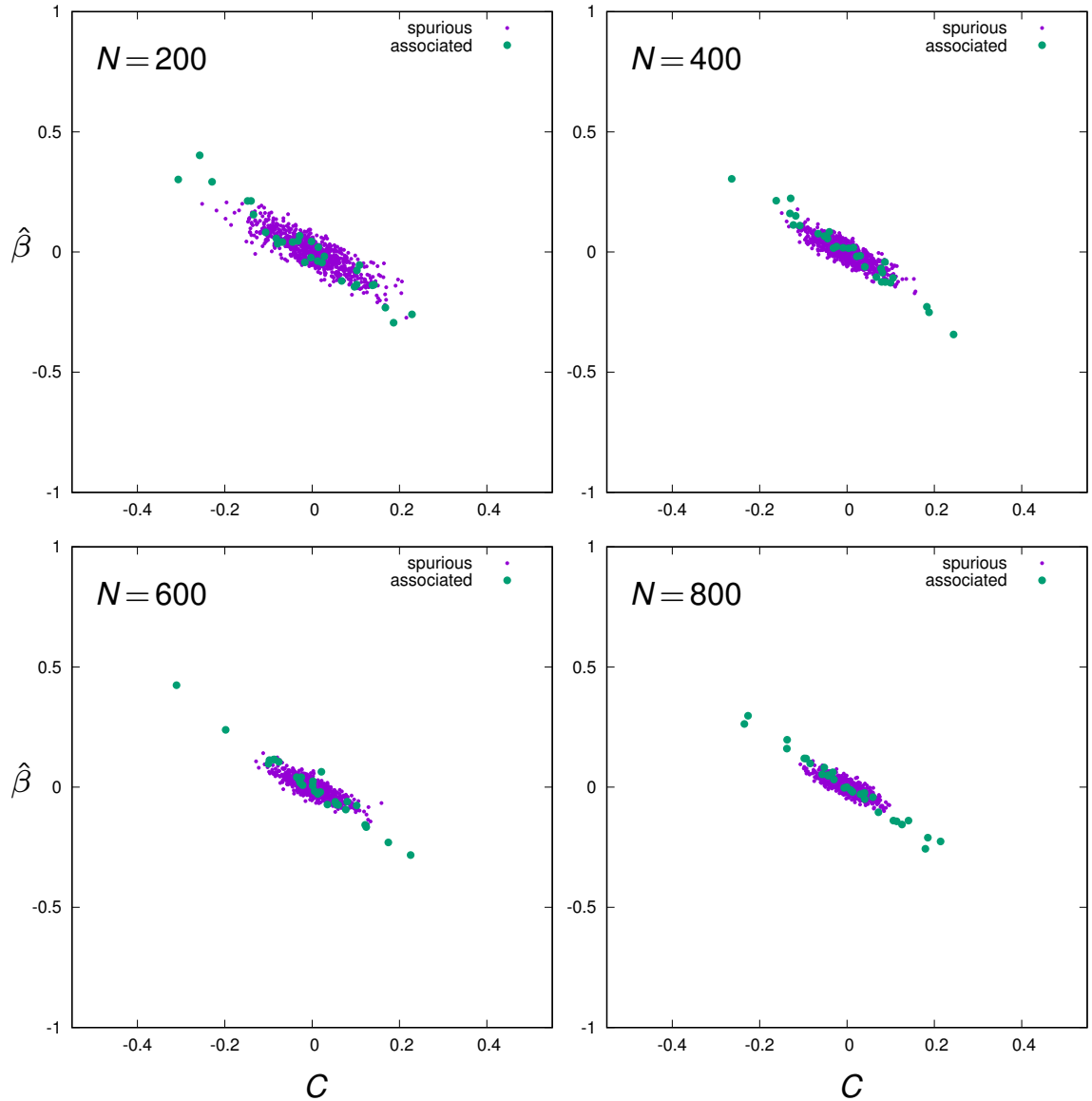
- $N \uparrow$: prob of false positive associations $\downarrow$
- $p \uparrow$: prob of false positive associations $\uparrow$

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

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



synthetic survival data,
with $p = 800$



Intuition for the problem

- information-theoretic interpretation of ML regression

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

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

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

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

- 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

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

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

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

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

  define:

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

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

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

- *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 $\sim 1/\gamma$

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

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

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

- application

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

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

Translation to Cox's model

$$p(t|\boldsymbol{z},\lambda,\beta) = \lambda(t)\, \mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}/\sqrt{p} - \Lambda(t)\exp(\boldsymbol{\beta}\cdot\boldsymbol{z}/\sqrt{p})}, \qquad \Lambda(t) = \int_0^t \mathrm{d}t'\ \lambda(t')$$

- since $\langle z_\mu \rangle$ and $\langle z_\mu z_\nu \rangle$ can be transformed away, definitions:

$$p(\boldsymbol{z}) = (2\pi)^{-d/2}\mathrm{e}^{-\frac{1}{2}\boldsymbol{z}^2}, \qquad p(t|\xi,\lambda) = \lambda(t)\mathrm{e}^{\xi - \Lambda(t)\exp(\xi)}$$
$$S^2 = (\beta^\star)^2/p, \qquad \lambda^\star = \lambda_0, \qquad \alpha = d/N$$

- Insert, work out, $N\to\infty$:

$$E(S,\lambda_0) = -\lim_{\gamma\to\infty}\lim_{n\to 0}\frac{1}{n}\frac{\partial}{\partial\gamma}\mathrm{extr}_{\{\boldsymbol{C},\lambda_1,\ldots,\lambda_n\}}\Big\{\frac{1}{2}\alpha n[1+\log(2\pi)] + \frac{1}{2}\alpha\log\mathrm{Det}(\boldsymbol{C}')$$
$$+\ \log\int\frac{\mathrm{d}\boldsymbol{y}\ \mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}}}{\sqrt{(2\pi)^{n+1}\mathrm{Det}\boldsymbol{C}}}\int\mathrm{d}t\ p(t|y_0,\lambda_0)\prod_{\alpha=1}^{n}\Big(\frac{p(t|y_\alpha,\lambda_\alpha)}{p(t|y_0,\lambda_0)}\Big)^\gamma\Big\}$$

$\boldsymbol{C}$: $(n+1)\times(n+1)$, $C_{ab} = \langle\beta^a\cdot\beta^b/p\rangle$, $a,b = 0\ldots n$

$\boldsymbol{C}'$: $n\times n$, $C'_{ab} = C_{ab} - C_{a0}C_{0b}/C_{00}^2$, $a,b = 1\ldots n$

Replica symmetric solution

If solution space connected:
saddle-point symmetric under *all* permutations of $\{1, \ldots, n\}$

$$\boldsymbol{C} = \begin{pmatrix} S^2 & c_0 & \cdots & \cdots & c_0 \\ c_0 & C & c & \cdots & c \\ \vdots & c & C & \cdots & c \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_0 & c & \cdots & c & C \end{pmatrix}, \qquad \lambda_\alpha(t) = \lambda(t) \quad \forall \alpha = 1 \ldots n$$

interpretation:

$$c_0 = \lim_{p\to\infty} \frac{1}{p} \boldsymbol{\beta}^\star \cdot \langle\langle \boldsymbol{\beta} \rangle\rangle_{\mathcal{D}}, \qquad c = \lim_{p\to\infty} \frac{1}{p} \langle\langle \boldsymbol{\beta} \rangle^2 \rangle_{\mathcal{D}}, \qquad C = \lim_{p\to\infty} \frac{1}{p} \langle\langle \boldsymbol{\beta}^2 \rangle\rangle_{\mathcal{D}}$$

Insert into formulae,
diagonalise $\boldsymbol{C}$ and $\boldsymbol{C}'$,
manipulations, integrations,
take limits $n \to 0$ and $\gamma \to \infty$ ...

- 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

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

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

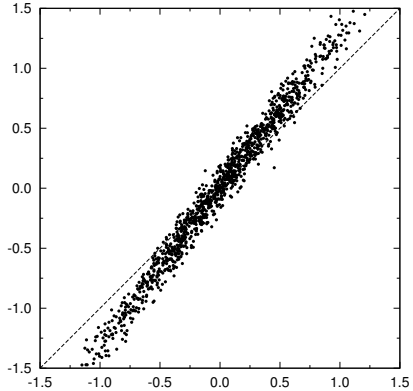
$W(x)$: Lambert function

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

- interpretation:

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

all we need for overfitting correction!

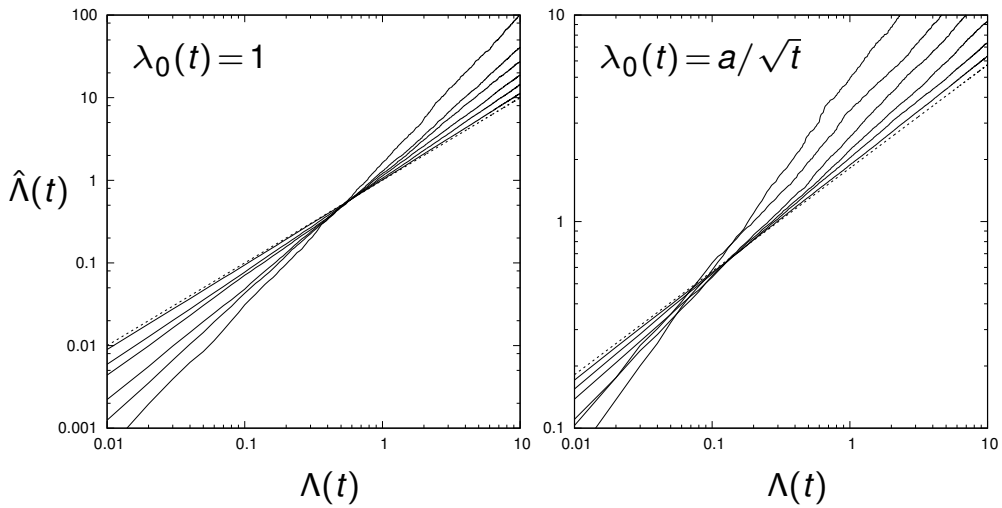


- challenge: eqn for $\Lambda(t)$

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

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

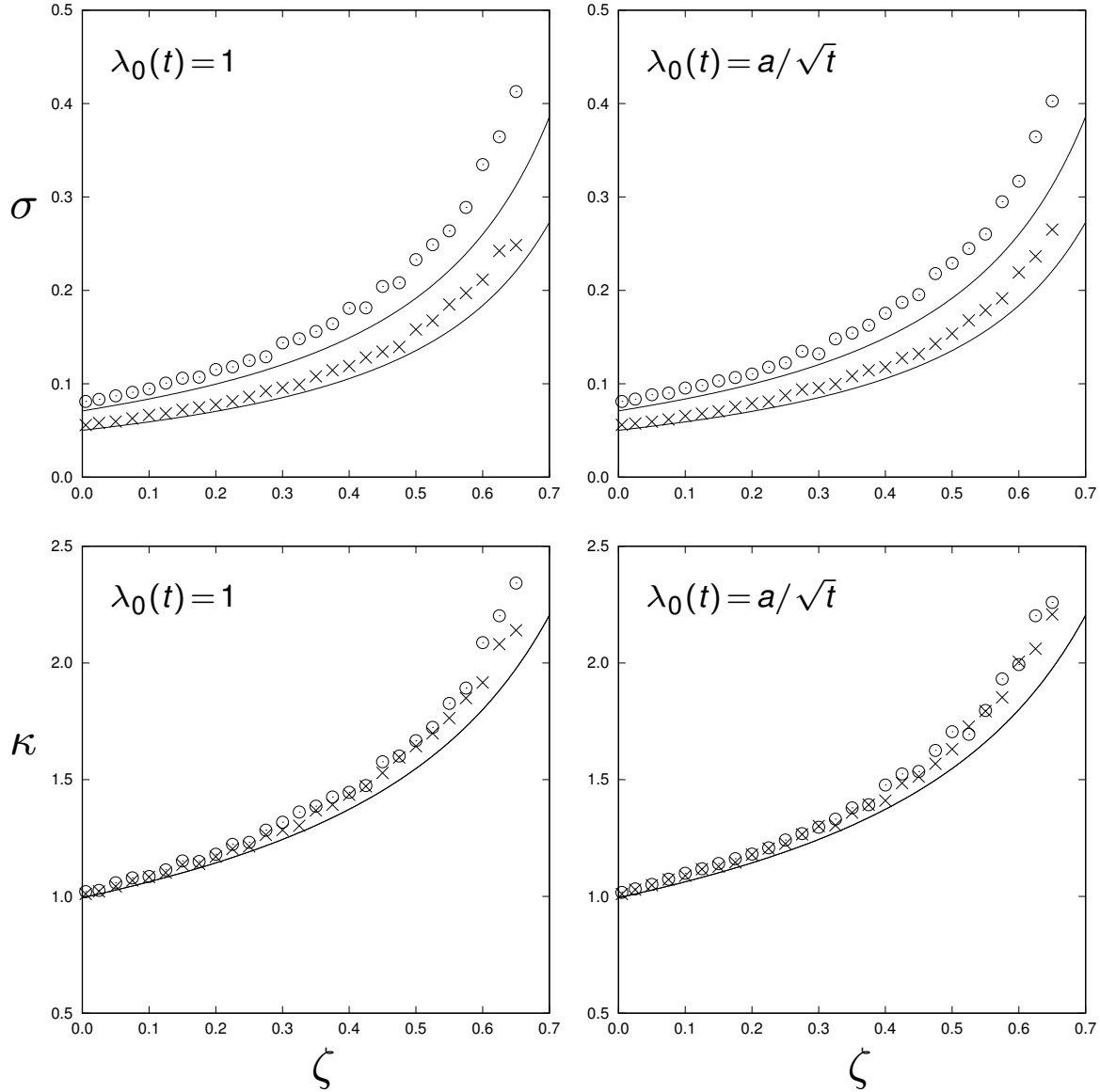
  $\Lambda^\star(t)$ drops out of equations!



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: $p(\boldsymbol{\beta}) \propto \exp(-\eta p \boldsymbol{\beta}^2)$

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 $\kappa = 1$

scalar order parameter eqns:

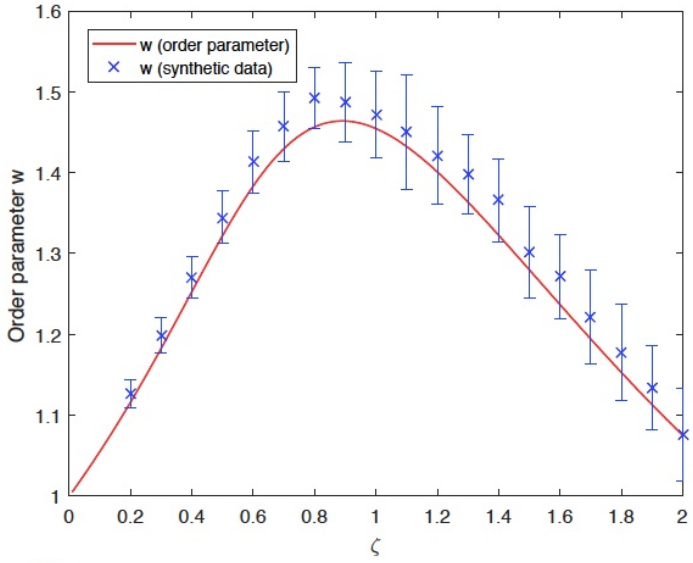
$$
\begin{aligned}
\zeta\tilde{f}\tilde{u}^4 &= -\int \mathrm{D}z\mathrm{D}y \int \mathrm{d}t\; p(t|S\langle a\rangle^{\frac{1}{2}}y,\lambda^\star)\Big[W\Big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy+vz}\Lambda(t)\Big)-\tilde{u}^2\Big]^2 \\
\zeta\tilde{g}\tilde{u}^2 &= \int \mathrm{D}z\mathrm{D}y \int \mathrm{d}t\; p(t|S\langle a\rangle^{\frac{1}{2}}y,\lambda^\star)\frac{W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy+vz}\Lambda(t)\big)}{1+W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy+vz}\Lambda(t)\big)} \\
0 &= \zeta w\Big[\langle a\rangle\Big\langle\frac{a^2}{2\eta+\tilde{g}a}\Big\rangle^{-1}-\tilde{g}\Big] \\
&\quad +\frac{1}{\tilde{u}^2}\int \mathrm{D}z\mathrm{D}y\; y\int \mathrm{d}t\; p(t|S\langle a\rangle^{\frac{1}{2}}y,\lambda^\star)W\Big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy+vz}\Lambda(t)\Big) \\
\tilde{u}^2 &= \Big\langle\frac{a}{2\eta+\tilde{g}a}\Big\rangle \\
v^2 &= w^2\Big[\langle a\rangle\Big\langle\frac{a^2}{2\eta+\tilde{g}a}\Big\rangle^{-2}\Big\langle\frac{a^3}{(2\eta+\tilde{g}a)^2}\Big\rangle-1\Big]-\tilde{f}\Big\langle\frac{a^2}{(2\eta+\tilde{g}a)^2}\Big\rangle
\end{aligned}
$$

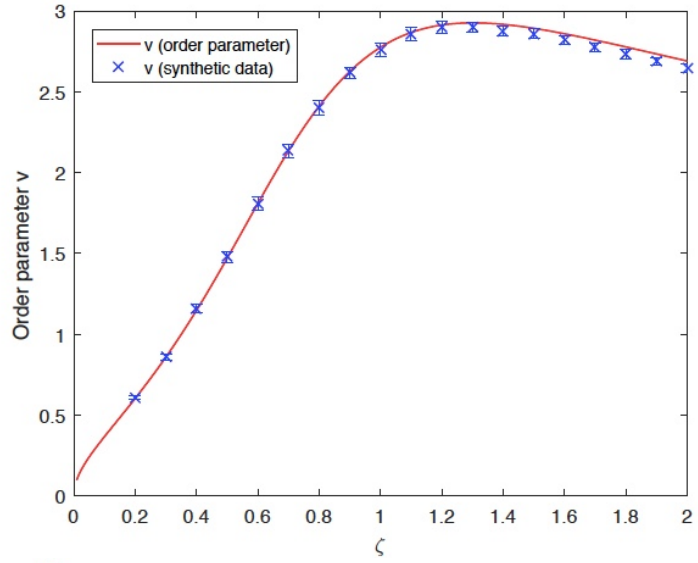
functional order parameter eqn:

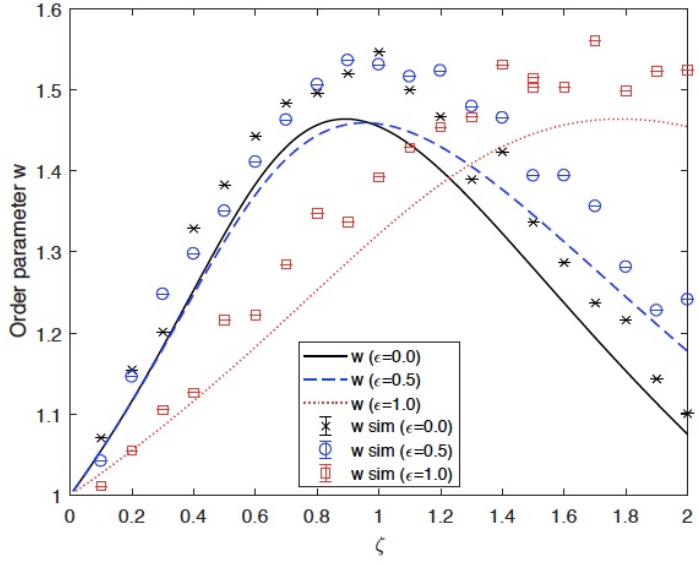
$$
\frac{p(t)}{\lambda(t)} = \int \mathrm{D}z\mathrm{D}y\int_t^\infty \frac{\mathrm{d}t'}{\tilde{u}^2\Lambda(t')}\; p(t'|S\langle a\rangle^{\frac{1}{2}}y,\lambda^\star)W\Big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy+vz}\Lambda(t')\Big)
$$

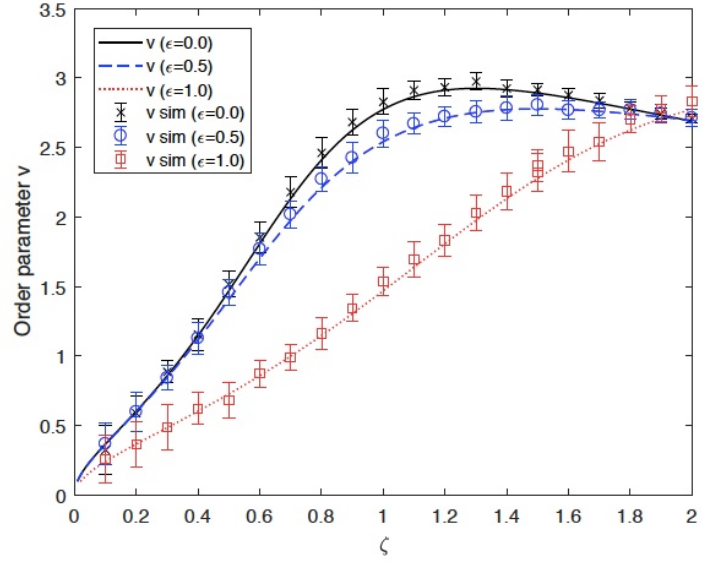
uncorrelated covariates

pairwise correlated covariates









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

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

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 $p(\beta) \propto \exp(-\eta p \beta^2)$ (small $\eta > 0$)
   result: $\hat{\beta}$ and $\hat{\Lambda}(t)$ (Breslow estimator)

3. calculate $\hat{\beta} \cdot \boldsymbol{A}\hat{\beta}$

4. solve coupled nonlinear eqns for (u, v, w, f, g, k, ρ),
   alongside $v^2 + w^2 = \hat{\beta} \cdot \boldsymbol{A}\hat{\beta}$
   (replaces unknown variable S)

5. calculate slope κ and noise amplitude σ

6. compute corrected estimators:

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

7. use σ to correct p-values

Generalized linear models

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

- *Logistic regression*

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

- *Ordinal class regression*

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

- *Latent class survival analysis*

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

- *Neural networks*

$$p(y|\boldsymbol{z},\{\boldsymbol{\beta}\},\boldsymbol{w},\boldsymbol{v}) = \mathrm{sgn}\left[\sum_{\ell=1}^{K} w_\ell\ \mathrm{sgn}\left[\boldsymbol{\beta}^\ell\cdot\boldsymbol{z}+v_\ell\right]+v_0\right], \qquad y=\pm 1$$

- .....

$$K=1: \qquad p(s|\boldsymbol{z},\boldsymbol{\theta})=p(s|\boldsymbol{\beta}\cdot\boldsymbol{z},\theta), \qquad p(\boldsymbol{\beta})\propto\exp(-\eta p\boldsymbol{\beta}^2)$$

replica analysis (RS):

order pars $(\tilde{f},\tilde{g},\tilde{u},v,w,\theta)$

$$\begin{aligned}
\Big\langle \frac{a}{2\eta+\tilde{g}a}\Big\rangle &= \tilde{u}^2 \\
w^2\Big[\langle a\rangle\Big\langle\frac{a^2}{2\eta+\tilde{g}a}\Big\rangle^{-2}\Big\langle\frac{a^3}{(2\eta+\tilde{g}a)^2}\Big\rangle-1\Big]-\tilde{f}\Big\langle\frac{a^2}{(2\eta+\tilde{g}a)^2}\Big\rangle &= v^2 \\
\int \mathrm{D}y\mathrm{D}z\ \langle[\xi(wy+vz,\tilde{u},s,\theta)-wy-vz]^2\rangle_s &= -\zeta\tilde{f}\tilde{u}^4 \\
\int \mathrm{D}y\mathrm{D}z\ \langle(\partial_1\xi)(wy+vz,\tilde{u},s,\theta)\rangle_s &= 1-\zeta\tilde{g}\tilde{u}^2 \\
\int \mathrm{D}y\mathrm{D}z\ \Big\langle\xi(wy+vz,\tilde{u},s,\theta)\frac{\partial\log p(s|S\langle a\rangle^{\frac{1}{2}}y,\theta^\star)}{\partial y}\Big\rangle_s &= \zeta w\tilde{u}^2\langle a\rangle\Big\langle\frac{a^2}{2\eta+\tilde{g}a}\Big\rangle^{-1} \\
\int \mathrm{D}y\mathrm{D}z\ \Big\langle\frac{\partial\log p(s|\xi(wy+vz,\tilde{u},s,\theta),\theta)}{\partial\theta}\Big\rangle_s &= 0
\end{aligned}$$

with

$$\langle F(s)\rangle_s=\int \mathrm{d}s\ p(s|S\langle a\rangle^{\frac{1}{2}}y,\theta^\star)F(s) \quad or \quad \sum_s p(s|S\langle a\rangle^{\frac{1}{2}}y,\theta^\star)F(s)$$

$$\xi(\mu,\sigma,s,\theta)=\mathrm{argmax}_{\xi\in\mathbb{R}}\Big[\log p(s|\xi,\theta)-\frac{1}{2}(\xi-\mu)^2/\sigma^2\Big]$$

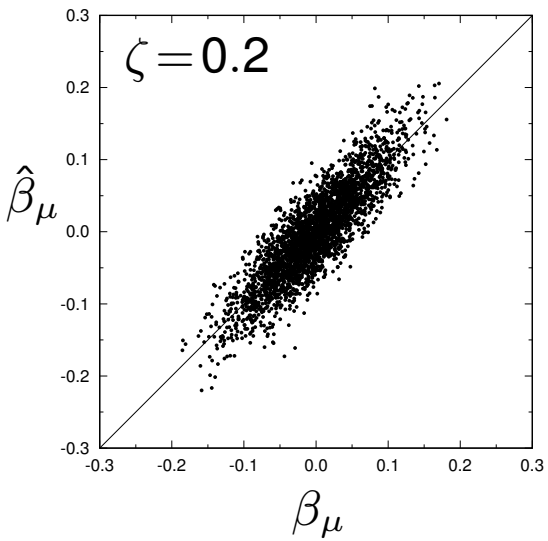
Test 1

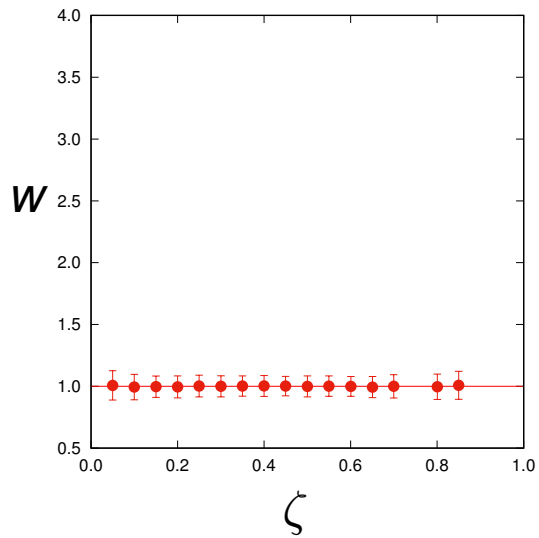
- *Linear regression* $s \in \mathbb{R}, \quad p(s|\boldsymbol{\beta}\cdot\boldsymbol{z},\beta_0) = \mathrm{e}^{-\frac{1}{2}(s-\boldsymbol{\beta}\cdot\boldsymbol{z}-\beta_0)^2/\Sigma^2}/\Sigma\sqrt{2\pi}$

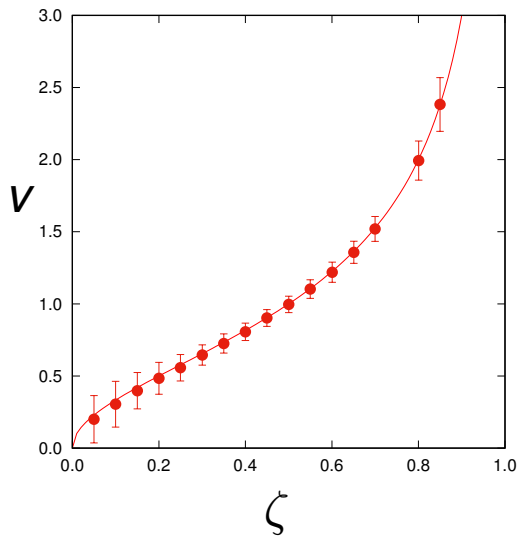
prediction for ML ($\eta = 0$):

$$\beta_0 = \beta_0^\star, \qquad w = S\langle a\rangle^{\frac{1}{2}}, \qquad \Sigma = \Sigma^\star\sqrt{1-\zeta} \quad \checkmark$$

$$\tilde{u} = \Sigma^\star\sqrt{\zeta}, \qquad v = \Sigma^\star\Big(\frac{\zeta}{1-\zeta}\Big)^{\frac{1}{2}}$$







$Np = 400,000$

no overfitting-induced bias,
but under-estimation of model noise

Test 2

- *Cox model*

$$t \geq 0, \quad p(t|\boldsymbol{\beta}\cdot\boldsymbol{z},\lambda) = \lambda(t)\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z} - \exp(\boldsymbol{\beta}\cdot\boldsymbol{z})\int_0^t \mathrm{d}t' \ \lambda(t')}$$

recover fully all previous

MAP/ML overfitting equations ✓

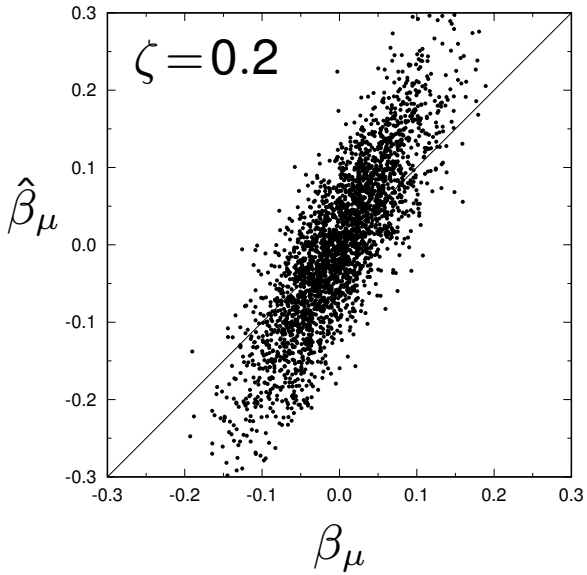
Application to logistic regression

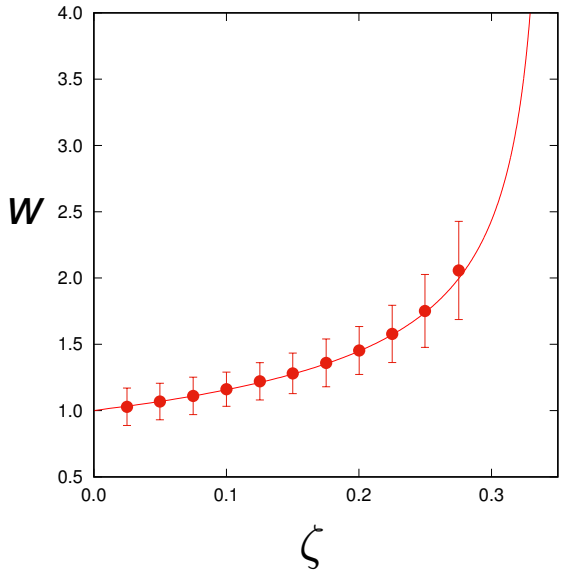
$$s = \pm 1, \qquad p(s|\boldsymbol{\beta}\cdot\boldsymbol{z},\beta_0) = \frac{\mathrm{e}^{s(\boldsymbol{\beta}\cdot\boldsymbol{z}+\beta_0)}}{2\cosh(\boldsymbol{\beta}\cdot\boldsymbol{z}+\beta_0)}$$

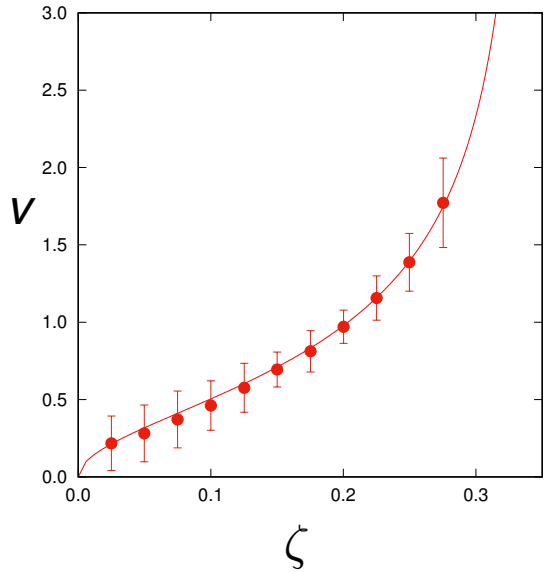
$$\begin{aligned}
\zeta v^2 &= \int \mathrm{D}y\mathrm{D}z \left\{ \frac{1}{2}\Big[1+\tanh(S\langle a\rangle^{\frac{1}{2}} y+\beta_0^\star)\Big]\Big[\tilde{x}(\beta_0+wy+vz,\tilde{u})-(\beta_0+wy+vz)\Big]^2 \right. \\
&\qquad \left. +\frac{1}{2}\Big[1-\tanh(S\langle a\rangle^{\frac{1}{2}} y+\beta_0^\star)\Big]\Big[\tilde{x}(-(\beta_0+wy+vz),\tilde{u})+(\beta_0+wy+vz)\Big]^2 \right\} \\
\zeta &= \int \mathrm{D}y\mathrm{D}z \left\{ \frac{1}{2}\Big[1+\tanh(S\langle a\rangle^{\frac{1}{2}} y+\beta_0^\star)\Big]\frac{\tilde{u}^2[1-\tanh^2(\tilde{x}(\beta_0+wy+vz,\tilde{u}))]}{1+\tilde{u}^2[1-\tanh^2(\tilde{x}(\beta_0+wy+vz,\tilde{u}))]} \right. \\
&\qquad \left. +\frac{1}{2}\Big[1-\tanh(S\langle a\rangle^{\frac{1}{2}} y+\beta_0^\star)\Big]\frac{\tilde{u}^2[1-\tanh^2(\tilde{x}(-(\beta_0+wy+vz),\tilde{u}))]}{1+\tilde{u}^2[1-\tanh^2(\tilde{x}(-(\beta_0+wy+vz),\tilde{u}))]} \right\} \\
\zeta w &= \frac{1}{2}S\langle a\rangle^{\frac{1}{2}} \int \mathrm{D}y\mathrm{D}z \Big[1-\tanh^2(S\langle a\rangle^{\frac{1}{2}} y+\beta_0^\star)\Big] \\
&\qquad \times\Big[\tilde{x}(\beta_0+wy+vz,\tilde{u})+\tilde{x}(-(\beta_0+wy+vz),\tilde{u})\Big] \\
\beta_0 &= \int \mathrm{D}y\mathrm{D}z \left\{ \frac{1}{2}\Big[1+\tanh(S\langle a\rangle^{\frac{1}{2}} y+\beta_0^\star)\Big]\tilde{x}(\beta_0+wy+vz,\tilde{u}) \right. \\
&\qquad \left. -\frac{1}{2}\Big[1-\tanh(S\langle a\rangle^{\frac{1}{2}} y+\beta_0^\star)\Big]\tilde{x}(-(\beta_0+wy+vz),\tilde{u}) \right\}
\end{aligned}$$

with

$$\tilde{x}(\eta,\sigma): \qquad \text{soln of} \quad \tanh(x) = 1 - (x-\eta)/\sigma^2$$







$Np = 400,000$

w: slope of $\{\hat{\beta}_\mu\}$ cloud

$v/\sqrt{p}$: width of $\{\hat{\beta}_\mu\}$ cloud

Overfitting in high-dimensional exponential inference models

observed data: $\boldsymbol{x}_i \in A^p$,
assumed model: $p(\boldsymbol{x}|\boldsymbol{\theta}^\star)$, $\boldsymbol{\theta}^\star \in \mathbb{R}^q$

- Exponential models:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{\mathrm{e}^{\boldsymbol{\theta}\cdot\boldsymbol{\omega}(\boldsymbol{x})/\sqrt{q}}}{|A|^p Z(\boldsymbol{\theta})}, \qquad Z(\boldsymbol{\theta}) = \frac{1}{|A|^p}\sum_{\boldsymbol{x}} \mathrm{e}^{\boldsymbol{\theta}\cdot\boldsymbol{\omega}(\boldsymbol{x})/\sqrt{q}}$$

  e.g: inference of
  – spin interactions from observed configurations
  – aminoacid contact maps from protein structures

- MAP inference, with Gaussian priors:

$$E_\gamma(\boldsymbol{\theta}^\star) = -\frac{\boldsymbol{\theta}^{\star 2}}{2N\sigma^2} - \frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{Nn}\log\int \mathrm{d}\boldsymbol{\theta}^1 \ldots \mathrm{d}\boldsymbol{\theta}^n \Big[\prod_{\alpha=1}^n \mathrm{e}^{-\frac{1}{2}\gamma\boldsymbol{\theta}^{\alpha 2}/\sigma^2}\Big]\, \mathrm{e}^{N\Psi(\{\boldsymbol{\theta}^\alpha\},\boldsymbol{\theta}^\star)}$$

$$\Psi(\{\boldsymbol{\theta}^\alpha\},\boldsymbol{\theta}^\star) = \log Z\Big(\boldsymbol{\theta}^\star + \gamma\sum_{\alpha=1}^n(\boldsymbol{\theta}^\alpha - \boldsymbol{\theta}^\star)\Big) - \gamma\sum_{\alpha=1}^n \log Z(\boldsymbol{\theta}^\alpha) - (1-\gamma n)\log Z(\boldsymbol{\theta}^\star)$$

Summary

- *Overfitting in Cox regression*

  bias in regression parameters: $\hat{\boldsymbol{\beta}} = \kappa \boldsymbol{\beta}^\star + \textit{noise}, \;\; \kappa > 1$
  bias in base hazard rates

- *Analytical approach based on the replica method*

  RS: closed equations for $\{u, v, w, \lambda(t)\}$
  variational approximation for $\lambda(t)$
  correct predictions for slope κ and noise σ

- *L2-regularized Cox regression*

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

- *Generalizations of the theory*

  arbitrary generalized linear regression models
  arbitrary high-dimensional exponential models

Papers, talks, seminars

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

Coolen et al, J Phys. A 50 (2017)
Sheikh and Coolen, J. Phys. A 52 (2019)

Thanks to

collaborators: James Barrett, Alexander Mozeika, Pierre Paga, Conrad Perez-Vicente, Mansoor Sheikh

funding: MRC, CRUK, GSK

January 2020

King's College London → Radboud University