Theory of overfitting in Cox regression

ACC Coolen, King's College London

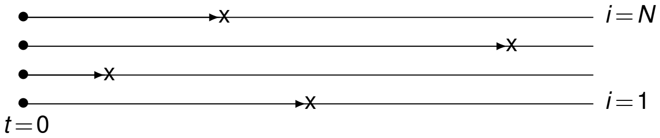
London, Sept 13th 2017

Regression for time-to-event data

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

  samples $(\boldsymbol{z}_i, t_i)$,
  drawn indep from $p(t, \boldsymbol{z})$

$$\boldsymbol{z}_i \in \mathbb{R}^d : \quad d \textit{ covariates } (\textit{measured at } t = 0)$$
$$t_i \in \mathbb{R}^+ : \quad \textit{failure time } (\textit{death}, \textit{ onset of disease}, \ldots)$$



- *Objective*

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

  1. *predict clinical outcome for individuals*
  2. *discover disease mechanisms*
  3. *design interventions (modifiable covariates)*

Proportional hazards regression
(DR Cox, 1972)

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

$$\textit{event time dist}: \quad p(t|\boldsymbol{z},\boldsymbol{\beta},\lambda) = -\frac{\mathrm{d}}{\mathrm{d}t}\exp\Big[-\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}}\int_0^t \mathrm{d}t'\ \lambda(t')\Big]$$

$$\textit{parameters}: \quad \boldsymbol{\beta} = (\beta_1,\ldots,\beta_d), \quad \lambda(t) \ \ t\geq 0$$

- Maximum Likelihood estimation

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

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

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

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

relatively simple and computationally painless,
extremely successful,
still the main tool of medical statisticians ...

Beyond the basic model ...

- *Fine tuning*
  - include left- right- or interval censoring (slightly different formula $p(\mathcal{D}|\boldsymbol{\beta}, \lambda)$)
  - consistent base hazard rate, such that $\int_0^\infty \mathrm{d}t\ \lambda(t) = \infty$ (ML subject to constraint $\int_0^\infty \mathrm{d}t\ \lambda(t) = R$, then $R \to \infty$)
- *Multiple risks*

  risk labels $r_i \in \{0, \ldots, R\}$,

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

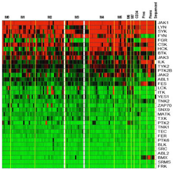
- *Frailty, random effects and latent class models*

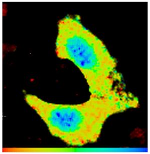
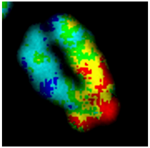
  (simple formulae only for special choices)

  $$p(t|\boldsymbol{z}, \boldsymbol{\beta}, \lambda) = -\frac{\mathrm{d}}{\mathrm{d}t} \sum_{\ell=1}^{L} w_\ell\ \exp\Big[-\mathrm{e}^{\boldsymbol{\beta}^\ell \cdot \boldsymbol{z}} \int_0^t \mathrm{d}t'\ \lambda^\ell(t')\Big]$$

What has changed since the 1970s?

- *Medical data have evolved*



- sheer *volume* ...
- *diversity* of data sources
  (clinical, genomic, biomarkers, health records, imaging, ...)
- *complexity* of experimental pipelines
  (confounders, batch effects, variability between centres, ...)
- *dimension* mismatch

  then: $\sim$500 samples, $\sim$10 covariates
  now: $\sim$1000 samples, $\sim 10^6$ covariates

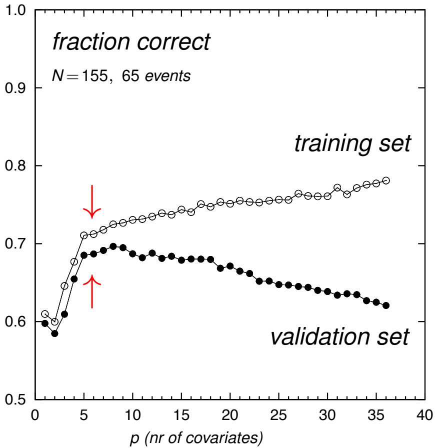
overfitting in
Cox regression

ML method ...

p-values, z-scores,
confidence intervals
don't measure overfitting!

rule of thumb: $p_{\max} = events/10$

- too optimistic ...
- must depend on $\boldsymbol{\beta}$ ...
- covariate correlations ...



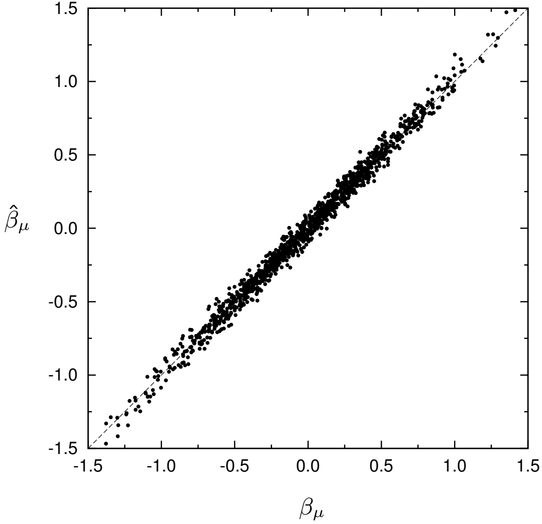
What happens in overfitting regime?

Can we predict the optimal point?

Analytical theory of overfitting in Cox regression?

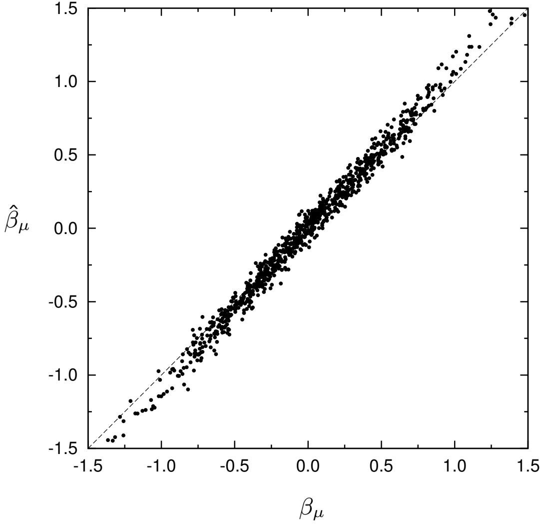
$N = 500$,
predicted versus true regression coefficients
(synthetic data, no censoring)

$p/N = 0.002$



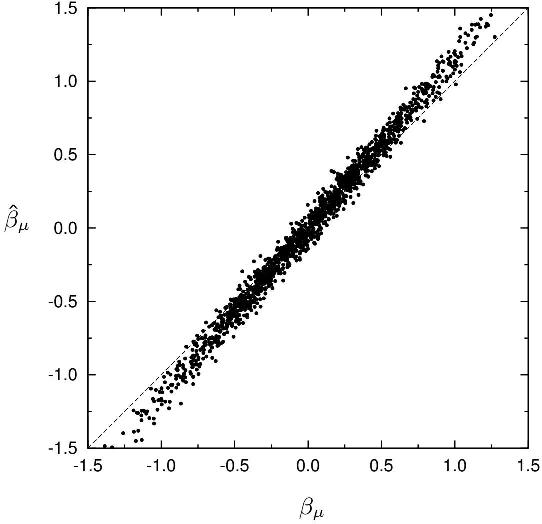
$N = 500$,
predicted versus true regression coefficients
(synthetic data, no censoring)

$p/N = 0.10$



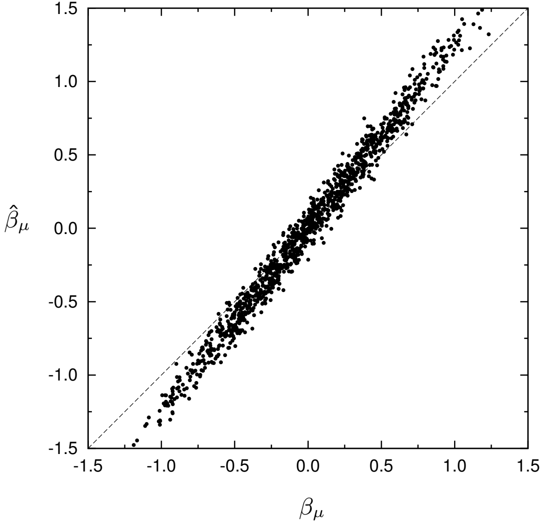
$N = 500$,
predicted versus true regression coefficients
(synthetic data, no censoring)

$p/N = 0.20$



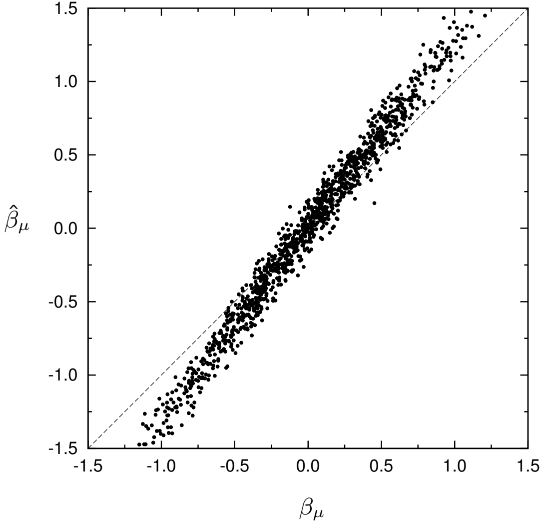
$N = 500$,
predicted versus true regression coefficients
(synthetic data, no censoring)

$p/N = 0.30$



$N = 500$,
predicted versus true regression coefficients
(synthetic data, no censoring)

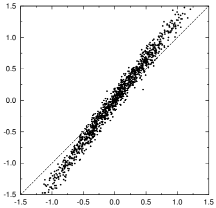
$p/N = 0.40$



Bad news

Overfitting *more dangerous*
than finite sample noise ...

we always inflate associations
(whether positive or negative)



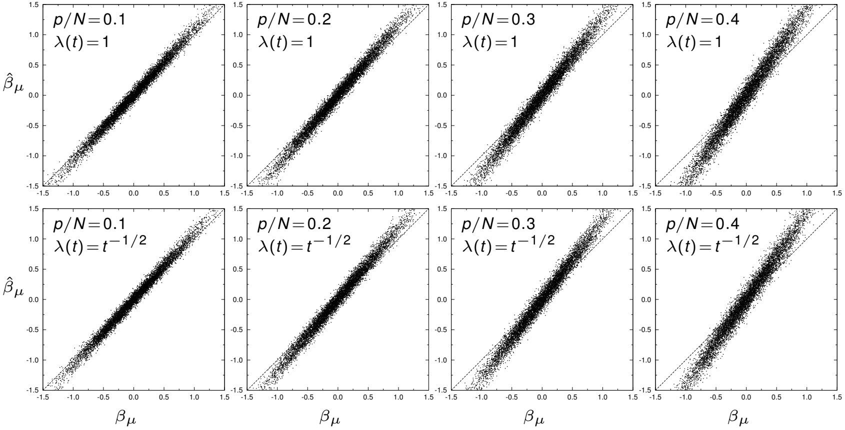
Good news

Unlike pure noise,
deterministic bias may be predictable ...

New possibilities, roadmap for research ...

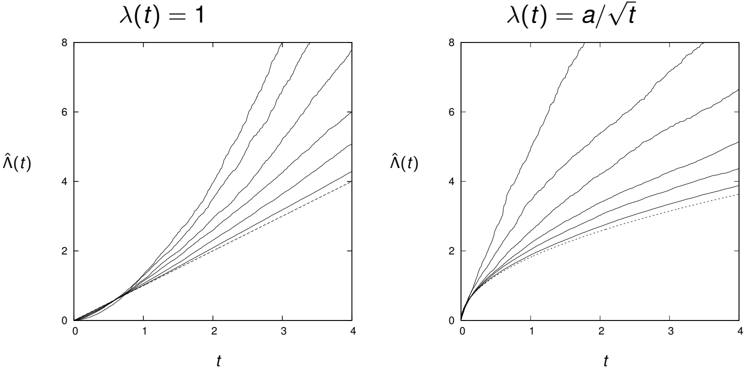
- Predict asymptotic impact of overfitting, in terms of
  – ratio p/N
  – correlations among covariates
  – true association strengths β

- Overfitting correction of Cox parameters
  – reliable regression at ratios $p/N \sim 0.5$ or more?



Association 'inflation' independent of true base hazard rate ...

$N = 400$,
Gaussian association pars,
$\langle \beta_\mu^2 \rangle = 0.25$



Base hazard rates underestimated for short times,
and over-estimated for large times ...

$p/N = 0.05, 0.15, 0.25, 0.35, 0.45, 0.55$
(lower to upper curves)

Gaussian association pars, $\langle \beta_\mu^2 \rangle = 0.25$,
$N = 400$, average event time $\langle t \rangle = 1$

Intuition for the problem ...

- *Overfitting in ML regression*

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

$$\boldsymbol{\theta}_{\rm ML} = \text{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}|\boldsymbol{\theta}) = \text{argmin}_{\boldsymbol{\theta}}\, D(\hat{p}||p_{\boldsymbol{\theta}})$$

$$\hat{p}(t,\mathbf{z}) = \frac{1}{N}\sum_i \delta(t-t_i)\delta(\mathbf{z}-\mathbf{z}_i), \qquad D(\hat{p}||p_{\boldsymbol{\theta}}) = \int \mathrm{d}t\mathrm{d}\mathbf{z}\; \hat{p}(t,\mathbf{z}) \log\Big[\frac{\hat{p}(t|\mathbf{z})}{p(t|\mathbf{z},\boldsymbol{\theta})}\Big]$$

  ML regression: move $p(t|\mathbf{z},\boldsymbol{\theta})$ towards $\hat{p}(t|\mathbf{z})$

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

  - fixed d: $\lim_{N\to\infty} \hat{p}(t,\mathbf{z}) = p(t,\mathbf{z}|\boldsymbol{\theta}^\star)$, so $\boldsymbol{\theta}_{\rm ML} = \boldsymbol{\theta}^\star$
  - $d = \mathcal{O}(N)$: $\lim_{N\to\infty} \hat{p}(t,\mathbf{z}) \neq p(t,\mathbf{z}|\boldsymbol{\theta}^\star)$ ...

- *Barrier to overfitting theory*

  want: study relation between $\boldsymbol{\theta}_{\rm ML}(\mathcal{D})$ and $\boldsymbol{\theta}^\star$, for $d = \mathcal{O}(N)$
  need: formula for $\boldsymbol{\theta}_{\rm ML}(\mathcal{D})$ ...

The basic ideas

Step 1 – define a suitable overfitting measure

- Let $\hat{p}_{\boldsymbol{\theta}^\star}$ be empirical distr of $(t, \boldsymbol{z})$, for data with true pars $\boldsymbol{\theta}^\star$

  note that

$$\boldsymbol{\theta}_{\rm ML} = \mathrm{argmin}_{\boldsymbol{\theta}}\, D(\hat{p}_{\boldsymbol{\theta}^\star}||p_{\boldsymbol{\theta}})$$

$$\boldsymbol{\theta} = \boldsymbol{\theta}^\star: \quad D(\hat{p}_{\boldsymbol{\theta}^\star}||p_{\boldsymbol{\theta}}) = D(\hat{p}_{\boldsymbol{\theta}^\star}||p_{\boldsymbol{\theta}^\star}) \quad \leftarrow \textit{not zero}!$$

  Define:

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

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

- *Typical behaviour*

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

□

Step 2 – eliminate minimisation over $\boldsymbol{\beta}$

- *Laplace identity (steepest descent)*

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

use in reverse:

$$\begin{aligned}
E(\boldsymbol{\theta}^\star) &= \Big\langle \min_{\boldsymbol{\theta}} \Big\{ \frac{1}{N} \sum_i \log \Big[\frac{p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta}^\star)}{p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta})} \Big] \Big\} \Big\rangle_{\mathcal{D}} \\
&= -\frac{1}{N} \Big\langle \max_{\boldsymbol{\theta}} \Big\{ \sum_i \log \Big[\frac{p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta})}{p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta}^\star)} \Big] \Big\} \Big\rangle_{\mathcal{D}} \\
&= -\lim_{\gamma\to\infty} \frac{1}{N} \Big\langle \frac{\partial}{\partial\gamma} \log \int \mathrm{d}\boldsymbol{\theta} \, \mathrm{e}^{\gamma \sum_i \log \left[p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta})/p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta}^\star) \right]} \Big\rangle_{\mathcal{D}} \\
&= -\lim_{\gamma\to\infty} \frac{1}{N} \frac{\partial}{\partial\gamma} \Big\langle \log \int \mathrm{d}\boldsymbol{\theta} \prod_{i=1}^{N} \Big[\frac{p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta})}{p(t_i|\boldsymbol{z}_i,\boldsymbol{\theta}^\star)} \Big]^\gamma \Big\rangle_{\mathcal{D}}
\end{aligned} \qquad \square$$

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

Step 3 – enable averaging over $\mathcal{D}$

- *Replica method* (tame the log ...)
$$\langle \log Z\rangle = \lim_{n\to 0}\frac{1}{n}\log\langle Z^n\rangle = \lim_{n\to 0}\frac{1}{n}\log\Big\langle \prod_{\alpha=1}^{n} Z\Big\rangle$$

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

- *Application*

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

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

□

Track record of the replica method
(Marc Kac, 1968)

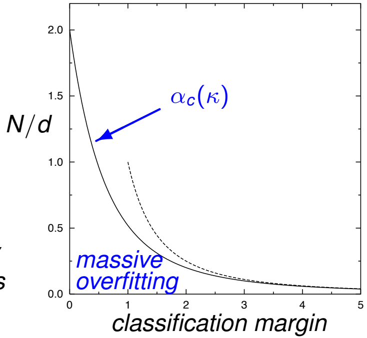
heterogeneous stochastic systems in physics, biology, computer science, economics, ...

- *disordered magnets (Sherrington & Kirkpatrick, 1975, Parisi, 1979)*
- *attractor neural networks (Amit, Gutfreund & Sompolinsky, 1985)*
- *solution space of binary classifiers (Gardner, 1988)*

since then:

satisfiability & optimisation problems, error-correcting codes, minority games, eigenvalue spectra of random graphs, machine learning, protein folding, immunology, compressed sensing, ...

Gardner theory for binary classifiers



Translation to Cox's model

$$p(t|\boldsymbol{z},\boldsymbol{\theta}) \quad\to\quad p(t|\boldsymbol{z},\lambda,\boldsymbol{\beta}) = \lambda(t)\,\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}/\sqrt{p}-\Lambda(t)\exp(\boldsymbol{\beta}\cdot\boldsymbol{z}/\sqrt{p})}$$

$$\Lambda(t) = \int_0^t \mathrm{d}t'\ \lambda(t')$$

- *Defns, short-hands*

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

$$S^2=\frac{1}{p}(\boldsymbol{\beta}^\star)^2, \qquad \lambda^\star=\lambda_0, \qquad \alpha=\frac{d}{N}$$

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

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

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

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

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 the limit $n \to 0$ ...
take the limit $\gamma \to \infty$ ...

Resulting asymptotic prediction

$$
\begin{aligned}
E(S,\lambda_0) &= \int \mathrm{d}t\; p(t)\log\left[\frac{\lambda_0(t)}{\lambda(t)}\right] - \\
&\quad (1+\tilde{u}^2)\Big[1-\frac{1}{\tilde{u}^2}\int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\; p(t|Sy_0,\lambda_0)\, W\Big(\tilde{u}^2 \mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\Big)\Big]
\end{aligned}
$$

$Dz = (2\pi)^{-1/2}\mathrm{e}^{-\frac{1}{2}z^2}\mathrm{d}z$,
$W(z)$: Lambert W-function,
$\tilde{u}, v, w, \lambda(t)$ to be solved from

$$
\begin{aligned}
\zeta v^2 &= \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\; p(t|Sy_0,\lambda_0)\Big[\tilde{u}^2 - W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\big)\Big]^2 \\
\zeta &= \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\; p(t|Sy_0,\lambda_0)\frac{W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\big)}{1+W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\big)} \\
0 &= \int \mathrm{D}z\mathrm{D}y_0\; y_0 \int \mathrm{d}t\; p(t|Sy_0,\lambda_0)\, W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\big) \\
\frac{p(t)}{\lambda(t)} &= \int \mathrm{D}z\mathrm{D}y_0 \int_t^{\infty} \mathrm{d}t'\; p(t'|Sy_0,\lambda_0)\frac{W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t')\big)}{\tilde{u}^2\Lambda(t')}
\end{aligned}
$$

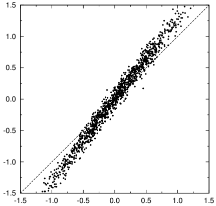
□

- interpretation:

$$v^2 = \frac{1}{p}\Big\{\langle \boldsymbol{\beta}^2\rangle_{\mathcal{D}} - \Big(\langle \boldsymbol{\beta}\rangle_{\mathcal{D}} \cdot \frac{\boldsymbol{\beta}^\star}{|\boldsymbol{\beta}^\star|}\Big)^2\Big\}, \qquad w = \frac{1}{\sqrt{p}}\langle \boldsymbol{\beta}\rangle_{\mathcal{D}} \cdot \frac{\boldsymbol{\beta}^\star}{|\boldsymbol{\beta}^\star|}$$

- link with data clouds:

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



- special limits for $\zeta = p/N$:

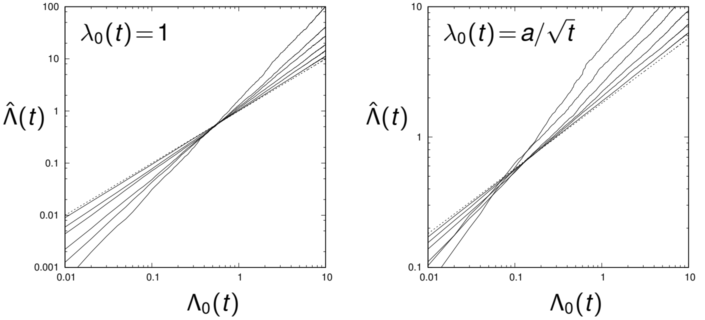
$$\zeta \to 0: \quad \textit{no overfitting}, \quad v \to 0, \;\; w \to S, \;\; \lambda(t) \to \lambda_0(t)$$
$$\zeta \to 1: \quad \textit{phase transition}, \quad v, w \to \infty$$

- main numerical challenge:

$$\frac{p(t)}{\lambda(t)} = \int \mathrm{D}z\mathrm{D}y_0 \int_t^\infty \mathrm{d}t'\; p(t'|Sy_0,\lambda_0)\frac{W\big(\tilde{u}^2 \mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t')\big)}{\tilde{u}^2\Lambda(t')}$$

$$t \gg 1: \quad \log\Lambda(t) = \rho\log\Lambda_0(t) + (1-\rho)\log(\log\Lambda_0(t)) + \ldots$$
$$\rho = \frac{w}{2S}\Big(1+\sqrt{1+4\tilde{u}^2/w^2}\Big)$$

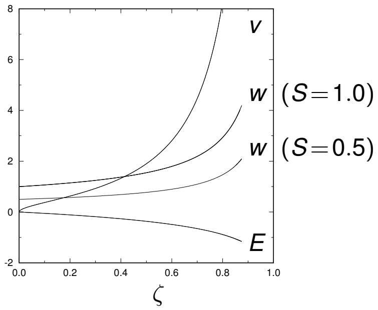
Variational approximation

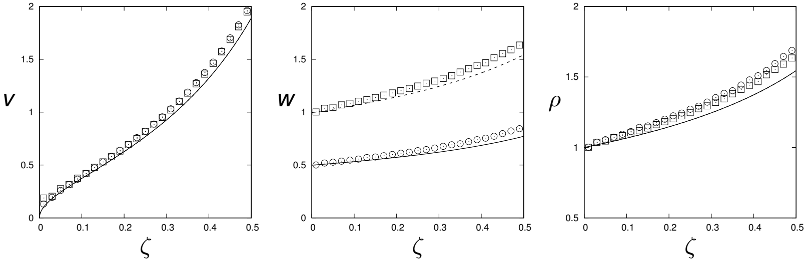


- (i) substitute ansatz $\Lambda(t) = k\Lambda_0^{\rho}(t)$ into extremization problem,
  (ii) work out variational eqns for $\tilde{u}, v, w, k, \rho$,
  (iii) solve numerically, gives $\rho = w/S$

- final theory:
  three coupled nonlinear eqns for $(w, \rho, \tilde{u})$

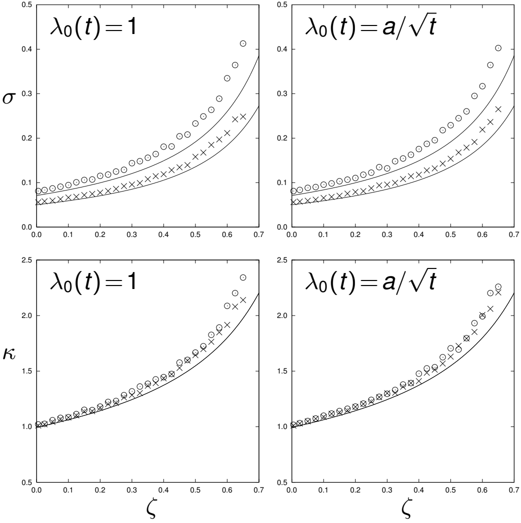
numerical solution
of variational eqns





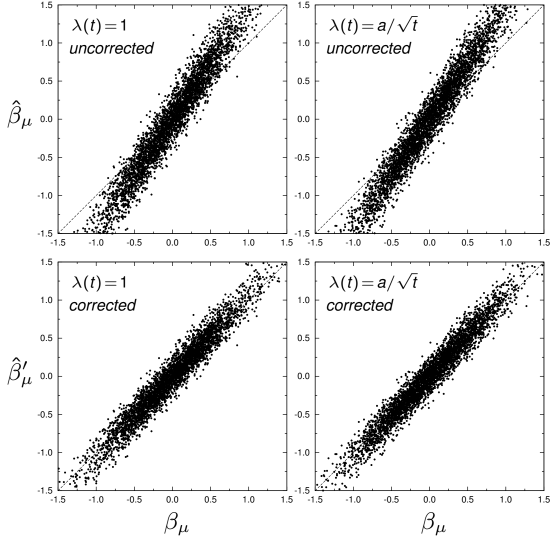
lines: predictions of variational theory

markers: simulations ($N = 200$), for $S = 0.5$ (o) and $S = 1$ (□)

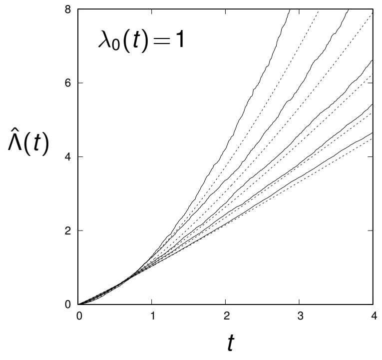


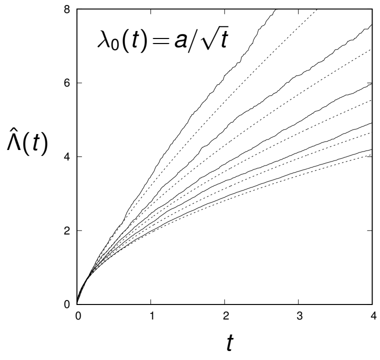
width σ and slope κ of data clouds
for $S = 0.5$ and $\langle t \rangle = 1$

lines: variational theory
o: $N = 200$
x: $N = 400$



overfitting correction of inflated association parameters using slope predicted by variational theory.
$N = 200$, $p = 80$, $\langle \beta_\mu^2 \rangle = 0.25$

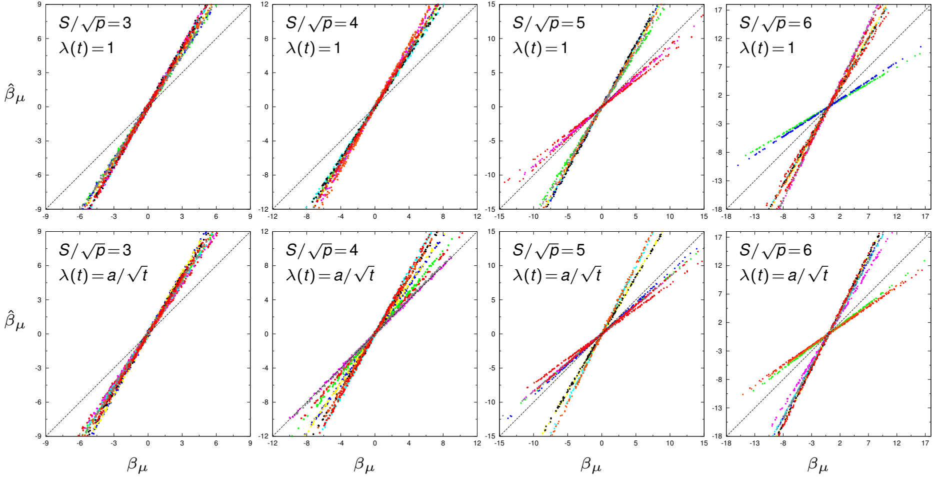




integrated base hazard rates
for $\zeta = 0.1, 0.2, 0.3, 0.4, 0.5$

all cases: $S = 0.5$ and $\langle t \rangle = 1$

dashed: variational theory
solid: simulations with $N = 400$



Large values of p/N and $\langle\beta_\mu^2\rangle$:
replica symmetry breaking
(disconnected solution spaces)

all cases: $N=500$, $\zeta=p/N=0.4$

Discussion

- Overfitting in Cox regression causes predictable bias

  (i) inflation of association parameters
  (ii) hazard rates: underestimated (t small), overestimated (t large)

- Analytical approach to model overfitting
  based on statistical mechanics (replica method)

- Replica symmetric theory:

  exact equations: $\{\tilde{u}, v, w, \lambda(t)\}$, nontrivial to solve numerically
  variational approximation: $\{\tilde{u}, w, \rho\}$, easy to solve numerically

- predictions of variational theory: quite good,
  reliable basis for overfitting corrections

- Next
  - Generalize to correlated covariates ✓✓
  - Include censoring ✓
  - Analysis of exact equations (no variational approx)
  - Associations for which $\sum_\mu \beta_\mu z_\mu$ is not Gaussian
  - Roll out overfitting correction protocols for Cox regression

Thank you!