Towards a theory of overfitting in proportional hazards regression for survival data

ACC Coolen

King's College London

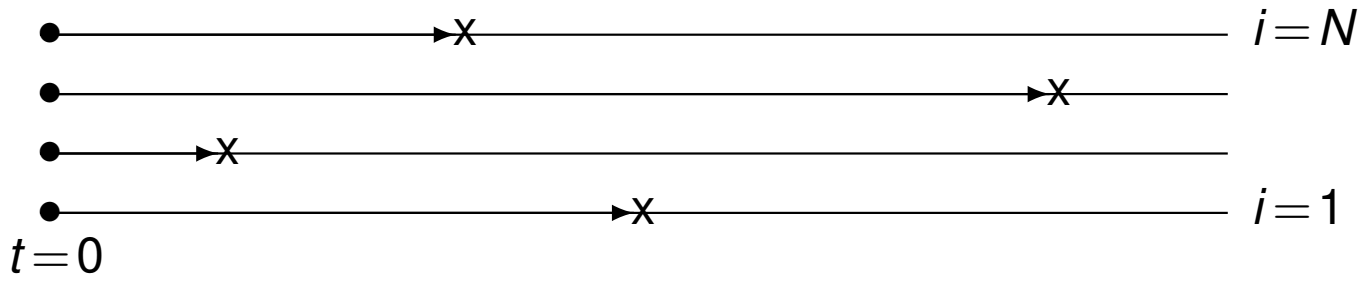
April 19th 2016

Regression for time-to-event data

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

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

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



- *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)

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

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

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

- Maximum Likelihood estimation

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

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

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

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

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

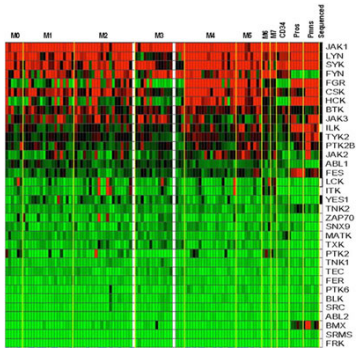
  $$\mathcal{D} = \{(\mathbf{z}_1, t_1, r_1), \ldots, (\mathbf{z}_N, t_N, r_N)\}$$
- *Frailty, random effects and latent class models*

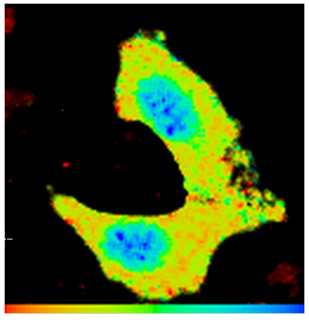
  (simple formulae only for special choices)

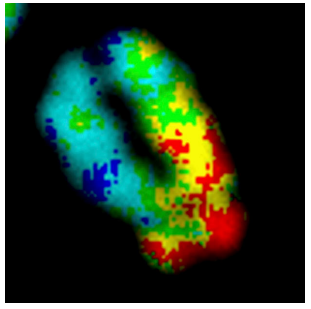
  $$p(t|\mathbf{z}, \boldsymbol{\beta}, \lambda) = -\frac{\mathrm{d}}{\mathrm{d}t} \sum_{\ell=1}^{L} w_\ell \ \exp\Big[-\mathrm{e}^{\boldsymbol{\beta}^\ell \cdot \mathbf{z}} \int_0^t \mathrm{d}t'\ \lambda^\ell(t')\Big]$$

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

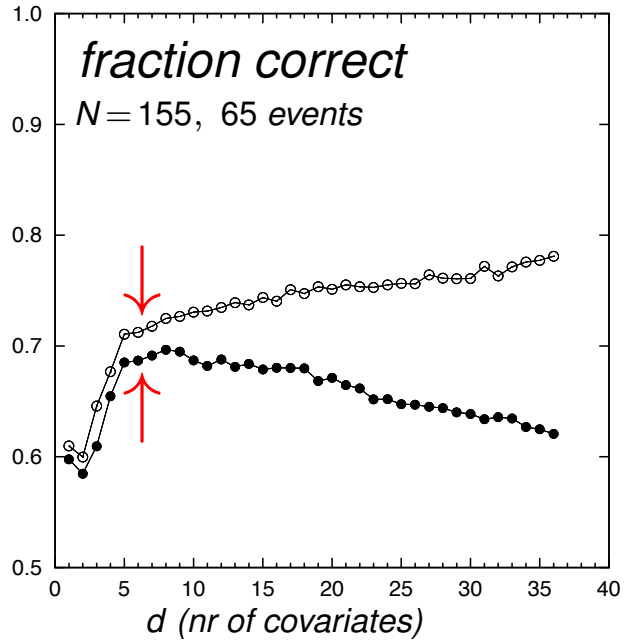
ML method ...

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

rule of thumb: $d_{\max} = N/10$

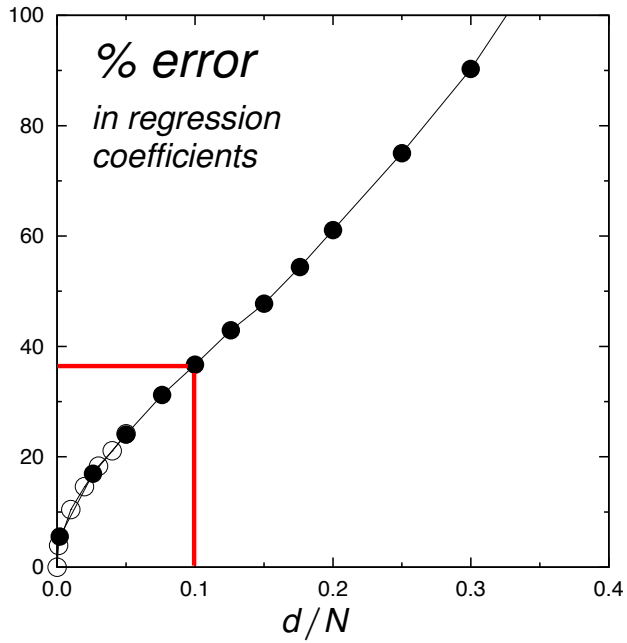
- too optimistic ...
- must depend on β ...
- must depend on cov correlations ...

Analytical theory of overfitting in Cox regression?



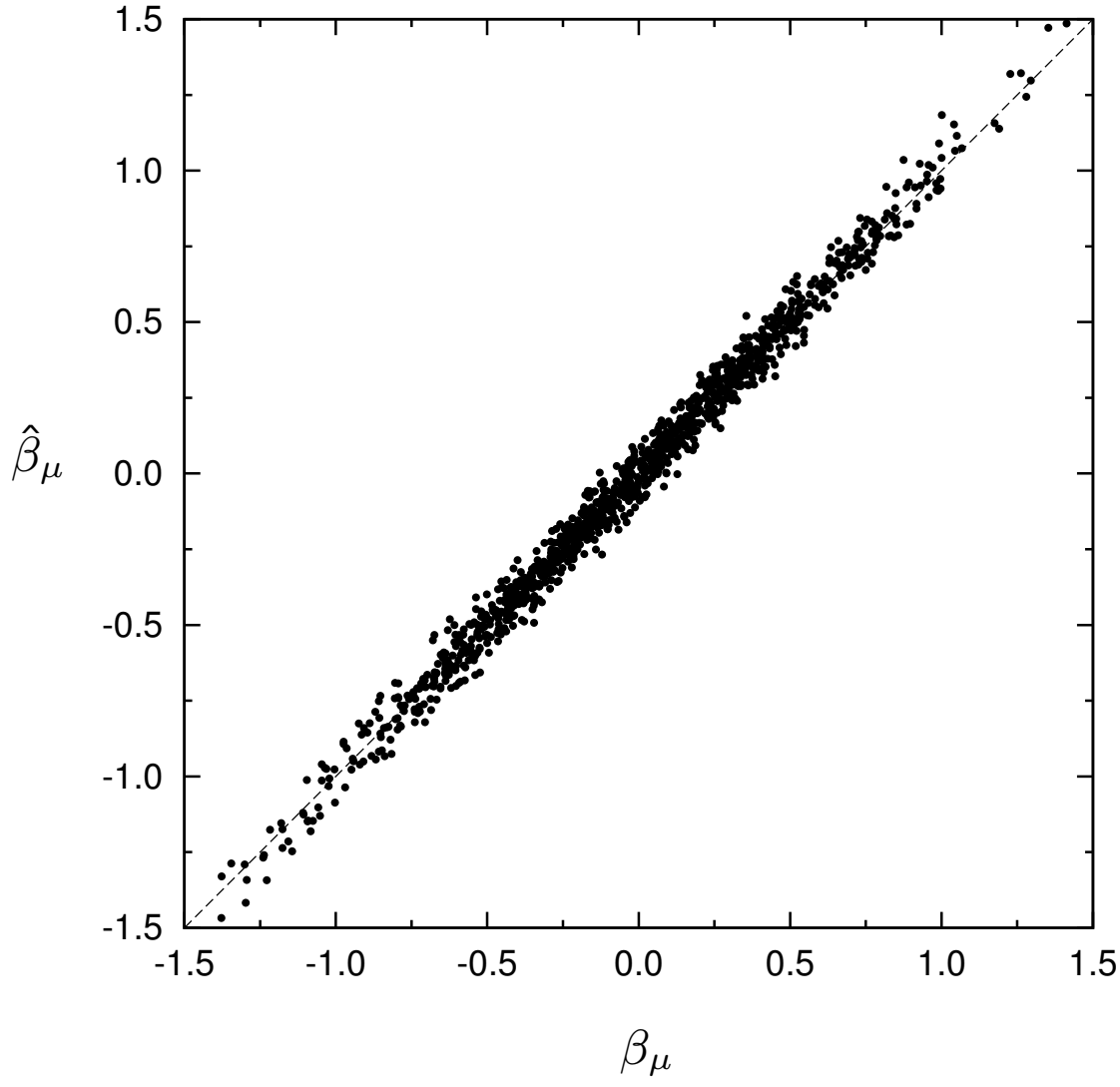
uncorrelated covars

o: $N = 1000$
●: $N = 500$



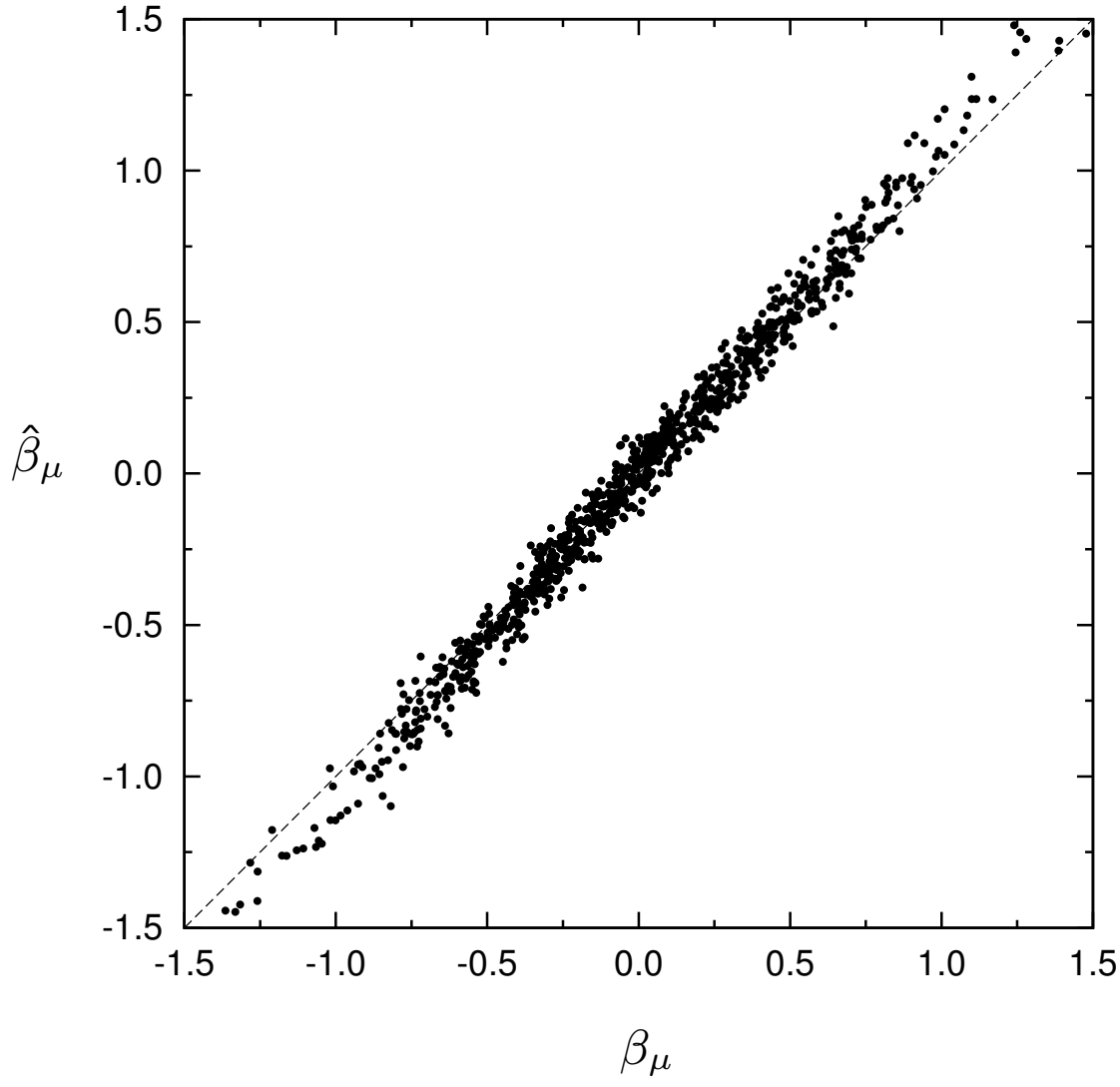
$N = 500$,
predicted versus true regression coefficients

$d/N = 0.002$



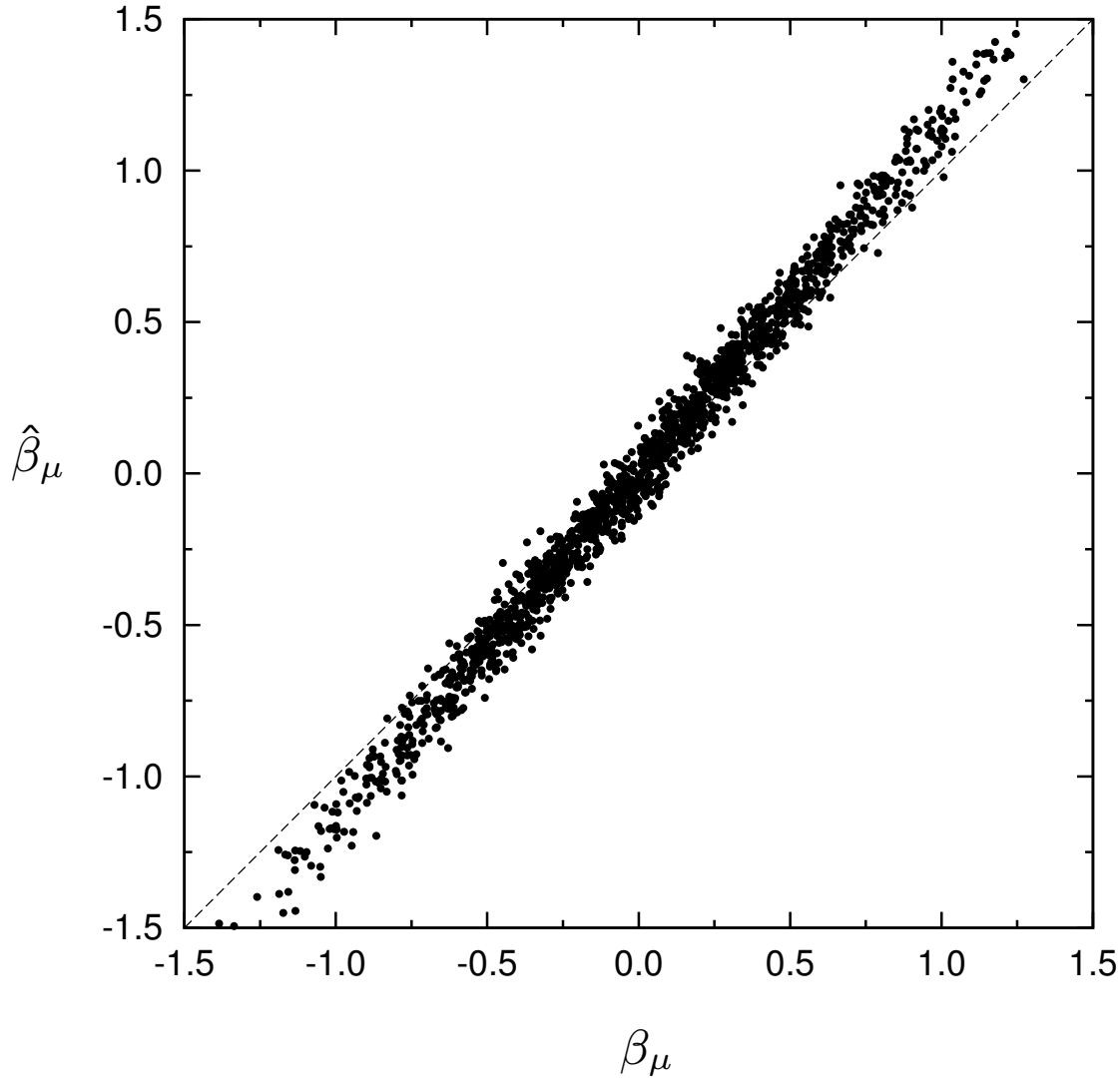
$N = 500$,
predicted versus true regression coefficients

$d/N = 0.10$



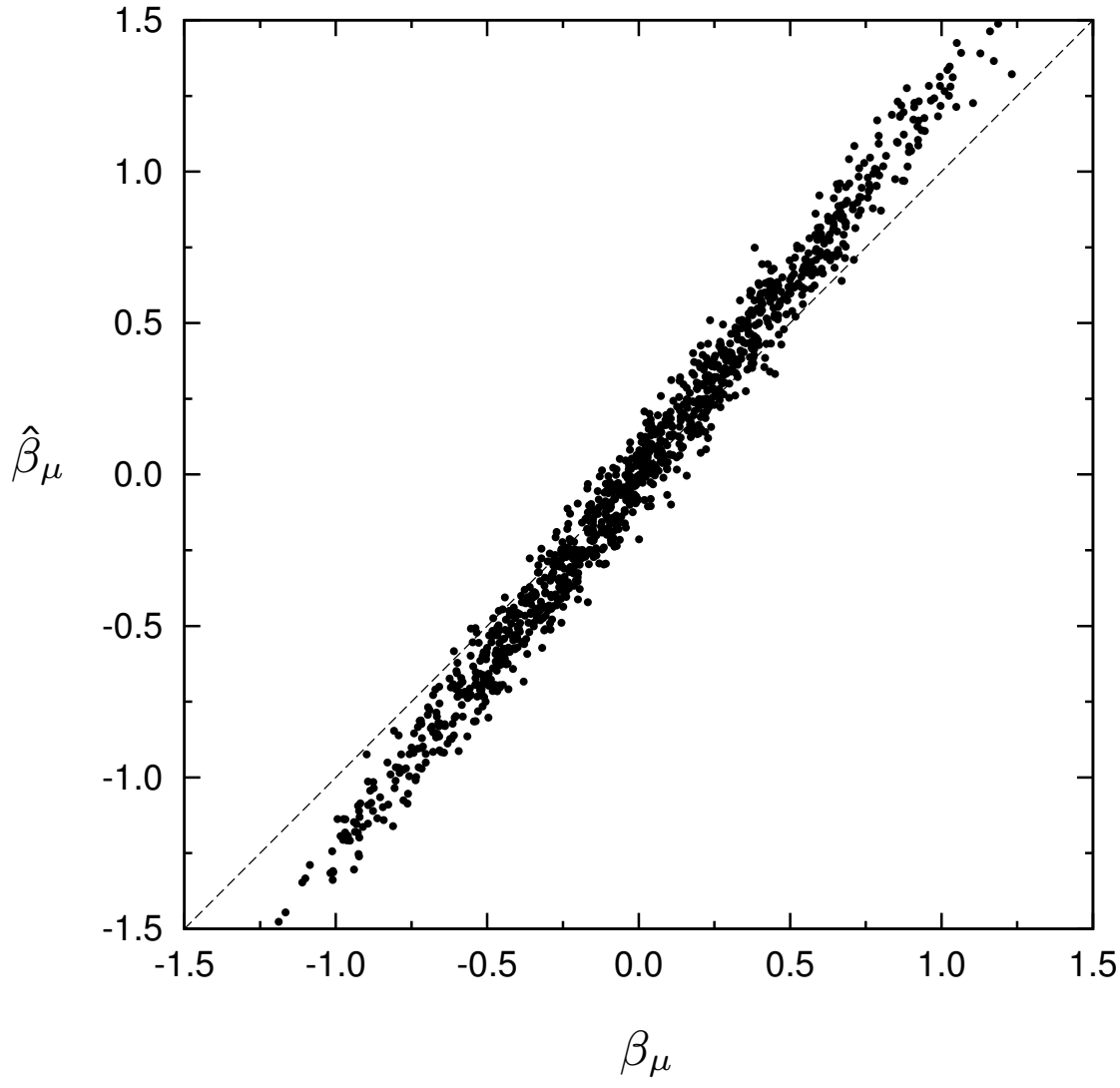
$N = 500$,
predicted versus true regression coefficients

$d/N = 0.20$



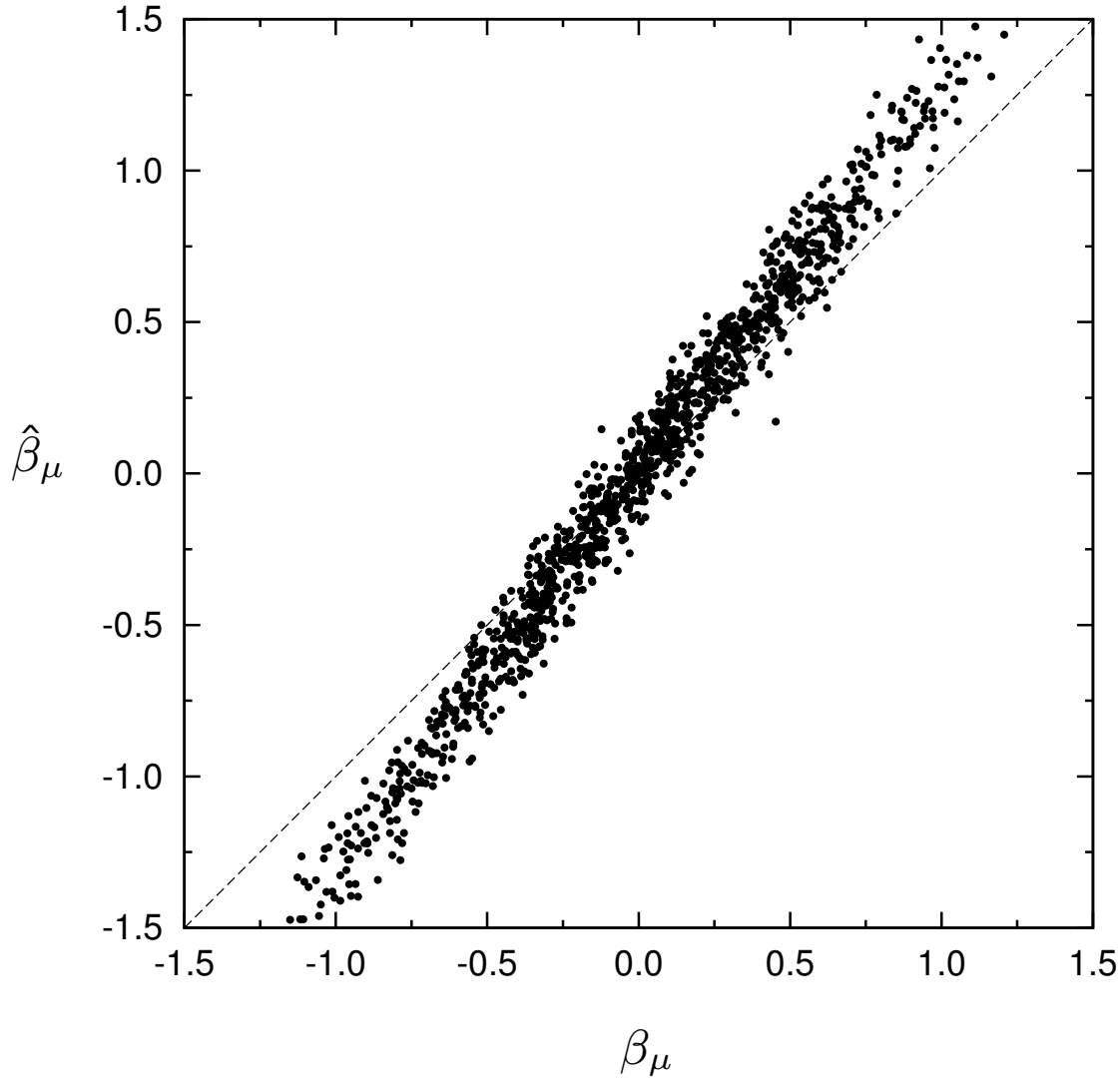
$N = 500$,
predicted versus true regression coefficients

$d/N = 0.30$



$N = 500$,
predicted versus true regression coefficients

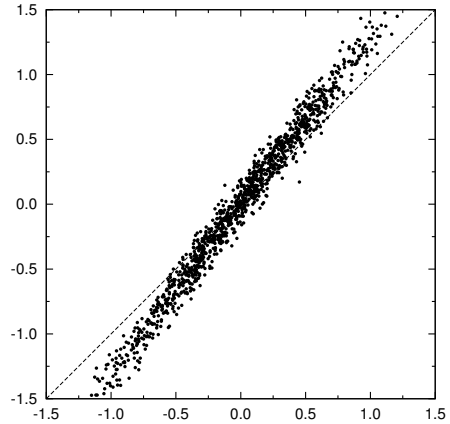
$d/N = 0.40$



Bad news

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

> *we always overestimate the*
> *strength of associations*
> *(whether positive or negative)*



Good news

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

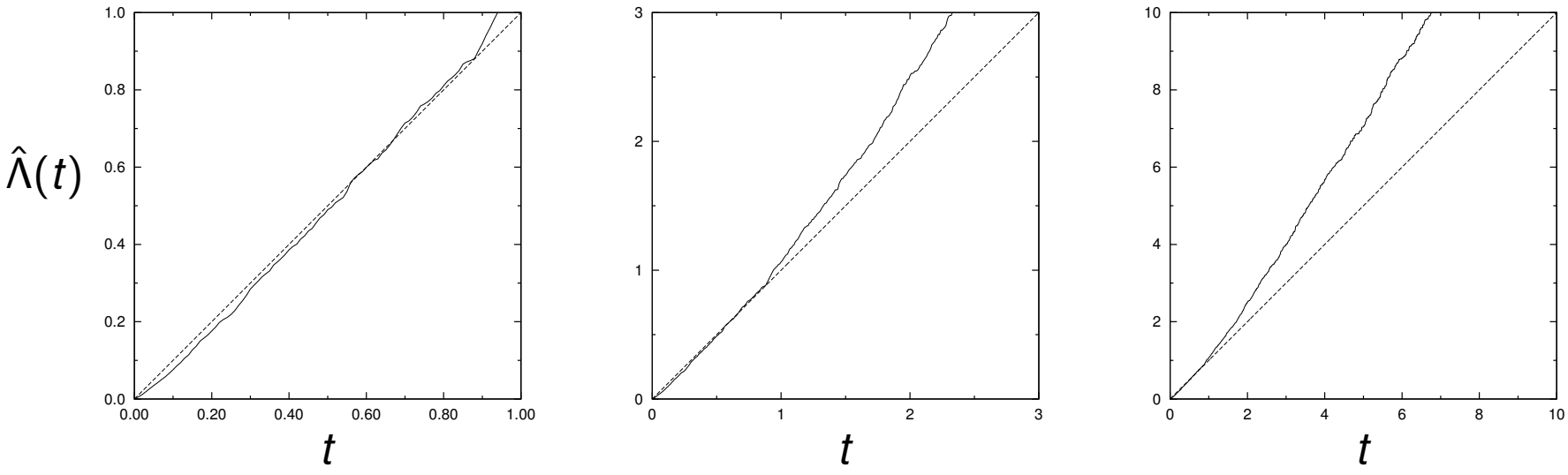
New possibilities, roadmap for research ...

- Predict asymptotic impact of overfitting, in terms of
  - ratio d/N
  - correlations among covariates
  - true association strengths β
- Overfitting correction of Cox parameters
  - reliable regression at ratios $d/N \sim 0.5$?
  - Bayesian Cox regression with unbiased estimates?

$N = 500, \quad d/N = 0.2$

inferred integrated base hazard rate
$\hat{\Lambda}(t) = \int_0^t \mathrm{d}t' \; \hat{\lambda}(t')$

dashed: true values



- deviations increase with ratio $\alpha = d/N$
- here: synthetic data with $\lambda(t) = 1$,
  so average event time for $\mathbf{z} = \mathbf{0}$ is 1

Intuition for the problem ...

- *Overfitting in ML regression*

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

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

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

  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 $\boldsymbol{\theta}_{\rm ML}(\mathcal{D})$ ...

  Cox regression: $\boldsymbol{\theta}_{\rm ML}(\mathcal{D})$ solved from transcendental eqn

The basic ideas

Step 1 – define a suitable overfitting measure

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

  note that

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

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

  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*

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

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

□

Step 2 – eliminate minimisation over β

- *Laplace identity*
  (steepest descent)

$$\lim_{\gamma\to\infty} \frac{1}{\gamma} \log \int \mathrm{d}x\, \mathrm{e}^{\gamma f(x)} \;=\; \max_x f(x)$$

use in reverse:

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

□

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*

$$\begin{aligned}
E(\boldsymbol{\theta}^\star) \;&=\; -\lim_{\gamma\to\infty} \frac{1}{\gamma N} \Big\langle \log \int \mathrm{d}\boldsymbol{\theta}\; \mathrm{e}^{\gamma \sum_i \log\left[p(t_i|\mathbf{z}_i,\boldsymbol{\theta})/p(t_i|\mathbf{z}_i,\boldsymbol{\theta}^\star)\right]} \Big\rangle_{\mathcal{D}} \\
&=\; -\lim_{\gamma\to\infty} \frac{1}{\gamma N} \lim_{n\to 0} \frac{1}{n} \log \Big\langle \Big[\int \mathrm{d}\boldsymbol{\theta}\; \mathrm{e}^{\gamma \sum_i \log\left[p(t_i|\mathbf{z}_i,\boldsymbol{\theta})/p(t_i|\mathbf{z}_i,\boldsymbol{\theta}^\star)\right]} \Big]^n \Big\rangle_{\mathcal{D}} \\
&=\; -\lim_{\gamma\to\infty} \lim_{n\to 0} \frac{1}{\gamma N n} \log \int \mathrm{d}\boldsymbol{\theta}^1 \ldots \mathrm{d}\boldsymbol{\theta}^n \Big\langle \mathrm{e}^{\gamma \sum_i \sum_{\alpha=1}^{n} \log\left[p(t_i|\mathbf{z}_i,\boldsymbol{\theta}^\alpha)/p(t_i|\mathbf{z}_i,\boldsymbol{\theta}^\star)\right]} \Big\rangle_{\mathcal{D}} \\
&=\; -\lim_{\gamma\to\infty} \lim_{n\to 0} \frac{1}{\gamma N n} \log \int \mathrm{d}\boldsymbol{\theta}^1 \ldots \mathrm{d}\boldsymbol{\theta}^n \Big[\int \mathrm{d}\mathbf{z}\mathrm{d}t\; p(\mathbf{z}) p(t|\mathbf{z},\boldsymbol{\theta}^\star) \prod_{\alpha=1}^{n} \Big(\frac{p(t|\mathbf{z},\boldsymbol{\theta}^\alpha)}{p(t|\mathbf{z},\boldsymbol{\theta}^\star)} \Big)^\gamma \Big]^N
\end{aligned}$$

□

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

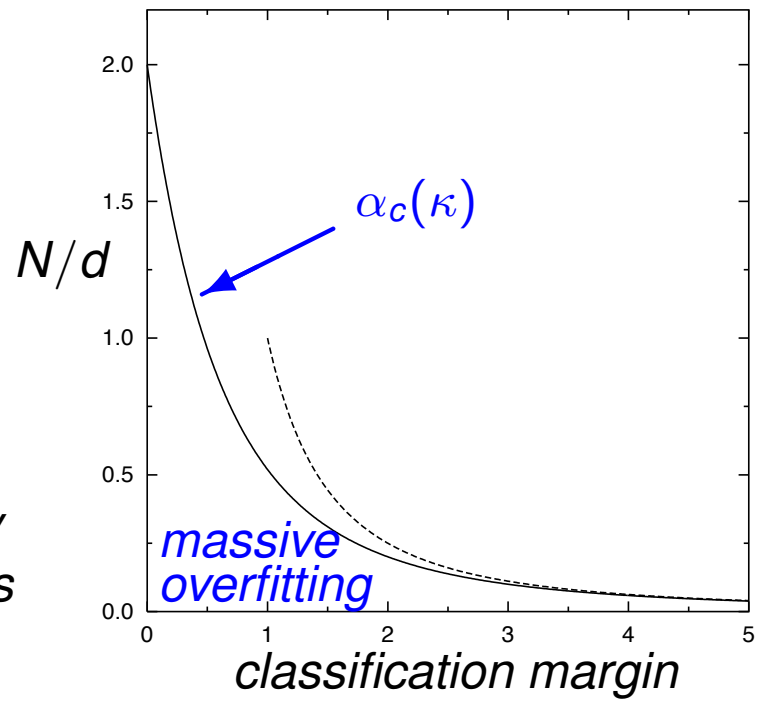
heterogeneous stochastic systems in physics,
biology, computer science, economics, ...

- *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

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

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

- *Defns, short-hands*

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

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

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

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

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

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

Replica symmetric solution

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

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

interpretation:

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

Insert into formulae,
diagonalise $\mathbf{C}$ and $\mathbf{C}'$, manipulations, integrations,
take the limit $n \to 0$ ...

The end game ...

$$E_{\rm RS}(S,\lambda_0) = -\lim_{\gamma\to\infty}\frac{1}{\gamma}{\rm extr}_{u,w;\lambda}\Big\{\alpha\log u$$
$$+\int\frac{{\rm d}y_0}{\sqrt{2\pi}}{\rm e}^{-\frac{1}{2}y_0^2}\int{\rm d}t\;p(t|Sy_0,\lambda_0)\log\int\frac{{\rm d}y}{\sqrt{2\pi}}{\rm e}^{-\frac{1}{2}y^2}\Big[\frac{p(t|uy+wy_0,\lambda)}{p(t|Sy_0,\lambda_0)}\Big]^\gamma\Big\}$$

□

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

- perfect fitting: $u\to 0$, $w\to S$, $\lambda(t)\to\lambda_0(t)$
  indeed gives: $E_{\rm RS}(S,\lambda_0)=0$
  provided $u\to 0$ slower than exponentially as $\gamma\to\infty$

- formula tricky, in view of
  (i) non-commutation of limits $\alpha\to 0$ vs $u\to 0$ vs $\gamma\to\infty$
  (ii) contains notorious integral
$$I(u,\kappa)=\int_{-\infty}^{\infty}{\rm d}y\;{\rm e}^{-\frac{1}{2}y^2-\kappa\exp(uy)}$$
  in Derrida 1981 for $u\to\infty$,
  here we need $\kappa\to\infty$ ...

Simple approximations

- *Approximate y-integral for finite $u > 0$*

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

$$E_{\mathrm{RS}}(S, \lambda_0) \approx \lim_{\gamma \to \infty} \mathrm{extr}_{\{u,w;\lambda\}} \Big\{ \frac{1-\alpha}{\gamma} \log u + \frac{w^2}{2\gamma u^2} - \int \mathrm{D}y \int \mathrm{d}t \; p(t|Sy, \lambda_0) \Big[\log\Big(\frac{\lambda(t)}{\lambda^\star(t)\Lambda(t)} \Big) - \frac{1}{\gamma u^2} \Big(\frac{1}{2} \log^2 \Lambda(t) + wy \log \Lambda(t) \Big) \Big] \Big\}$$

- *Extremise over w and u*

$$w = - \int \mathrm{D}y \; y \int \mathrm{d}t \; p(t|Sy, \lambda_0) \log \Lambda(t)$$

$$u^2 = \frac{1}{1-\alpha} \Big\{ \int \mathrm{D}y \int \mathrm{d}t \; p(t|Sy, \lambda_0) \log^2 \Lambda(t) - \Big[\int \mathrm{D}y \; y \int \mathrm{d}t \; p(t|Sy, \lambda_0) \log \Lambda(t) \Big]^2 \Big\}$$

$$E_{\mathrm{RS}}(S, \lambda_0) \approx \lim_{\gamma \to \infty} \mathrm{extr}_{\{\lambda\}} \Big\{ - \int \mathrm{D}y \int \mathrm{d}t \; p(t|Sy, \lambda_0) \log\Big(\frac{\lambda(t)}{\lambda^\star(t)\Lambda(t)} \Big) + \frac{1-\alpha}{2\gamma} \log \Big[\int \mathrm{D}y \int \mathrm{d}t \; p(t|Sy, \lambda_0) \log^2 \Lambda(t) - \Big[\int \mathrm{D}y \; y \int \mathrm{d}t \; p(t|Sy, \lambda_0) \log \Lambda(t) \Big]^2 \Big] \Big\}$$

- *Choose* $\lambda_0(t) = \lambda_0$

  Variational approx for inferred base hazard rate: $\lambda(t) = \lambda$

  results in

$$w = S, \qquad u^2 = \frac{\pi^2/6}{1-\alpha}, \qquad \lambda = \lambda_0 e^{C_E}$$

  C_E: Euler's constant, ~ 0.5772

  - everything scales sensibly with γ, limits can be taken
  - u diverges for $\alpha \to 1$ (confirmed by simulations)
  - $\lambda > \lambda_0$ (confirmed by simulations)

  but approx of course too crude ...

  - should have w and λ dependent on α ...
  - should have $\lim_{\alpha \to 0} u = 0$ ...

Discussion

- Overfitting in PH causes predictable bias in regression pars
- Analytical approach to model overfitting in PH regression
- Problem reduced to analysis of an integral
- Preliminary approximations give rough but sensible results

- Next:
  - Tackle remaining integral properly
  - Generalise to correlated Gaussian covariates
  - Predict overfitting correction factors
  - Stochastic minimisation at optimal γ, such that $E = 0$

Thank you!