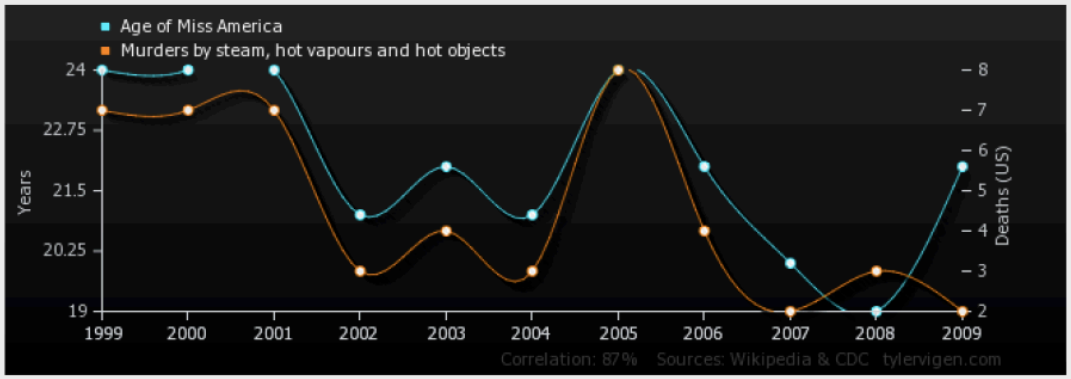
Replica analysis of overfitting in Cox regression

Statistical Physics Approaches to Systems Biology, Havana, Feb 2019

ACC Coolen

King's College London and Saddle Point Science

Age of Miss America
correlates with
Murders by steam, hot vapours and hot objects



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Age of Miss America *Years (Wikipedia)*	24	24	24	21	22	21	24	22	20	19	22
Murders by steam, hot vapours and hot objects *Deaths (US) (CDC)*	7	7	7	3	4	3	8	4	2	3	2

Correlation: 0.870127

Proportional hazards regression

(David Cox, 1972)

one risk,
no censoring:

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

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

- Maximum likelihood:

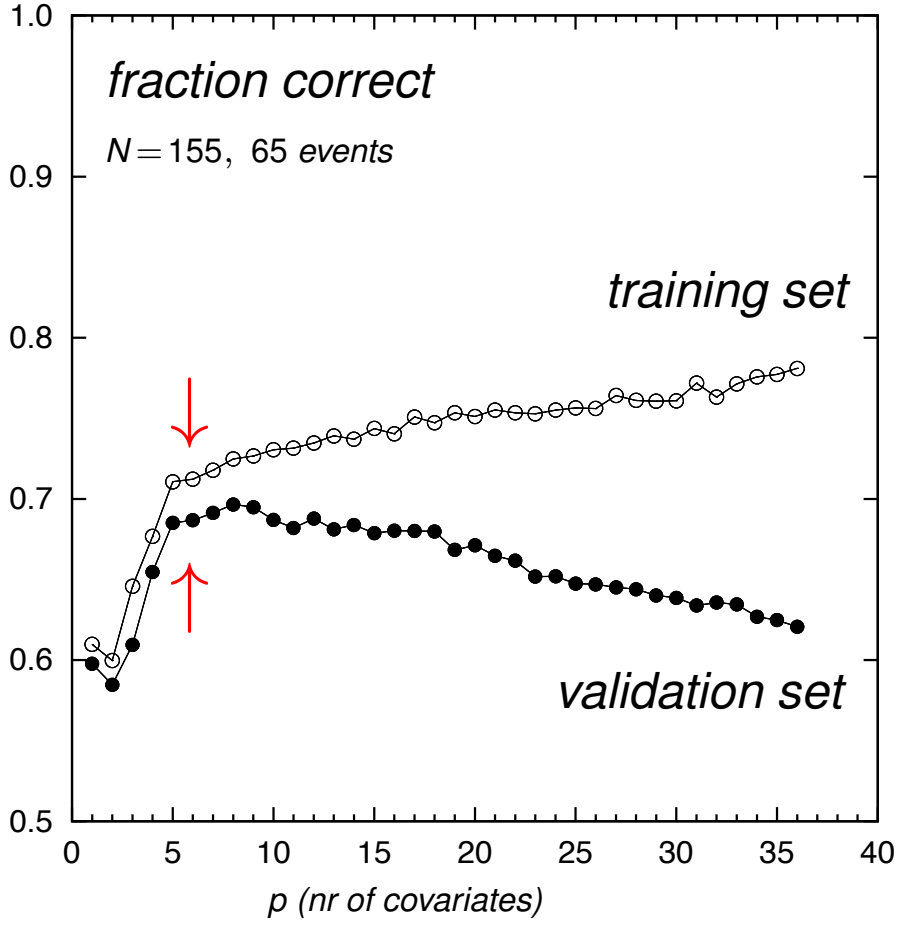
$$(\hat{\boldsymbol{\beta}}, \hat{\lambda}) \quad = \quad \mathrm{argmax}_{\boldsymbol{\beta},\lambda}\Big\{\frac{1}{N}\sum_i \log p(t_i|\boldsymbol{z}_i, \boldsymbol{\beta}, \lambda)\Big\}$$

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

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

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

overfitting in
Cox regression



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

- too optimistic?
- dependent on β?
- covariate correlations?

what happens in overfitting regime?

can we predict the optimal point?

analytical theory of overfitting in Cox regression?

theory simplification: choose $\langle z_\mu \rangle = 0$ and $\langle z_\mu z_\nu \rangle = \delta_{\mu\nu}$,
since average and correlations can be transformed away ...

- define

$$\langle f(\boldsymbol{z}) \rangle = \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{z}_i), \qquad A_{\mu\nu} = \langle z_\mu z_\nu \rangle - \langle z_\mu \rangle \langle z_\nu \rangle, \qquad \boldsymbol{z}_i = \langle \boldsymbol{z} \rangle + \boldsymbol{A}^{\frac{1}{2}} \tilde{\boldsymbol{z}}_i$$

- Cox regression:

$$\hat{\lambda}(t) = \mathrm{e}^{-\hat{\boldsymbol{\beta}} \cdot \langle \boldsymbol{z} \rangle} \frac{\sum_i \delta(t - t_i)}{\sum_i \theta(t_i - t) \mathrm{e}^{\tilde{\boldsymbol{\beta}} \cdot \tilde{\boldsymbol{z}}_i}}$$

$$\hat{\boldsymbol{\beta}} = \mathrm{argmax}_{\boldsymbol{\beta}} \sum_i \left\{ (\boldsymbol{A}^{\frac{1}{2}} \boldsymbol{\beta}) \cdot \tilde{\boldsymbol{z}}_i - \log \left[\sum_j \theta(t_j - t_i) \mathrm{e}^{(\boldsymbol{A}^{\frac{1}{2}} \boldsymbol{\beta}) \cdot \tilde{\boldsymbol{z}}_j} \right] \right\}$$

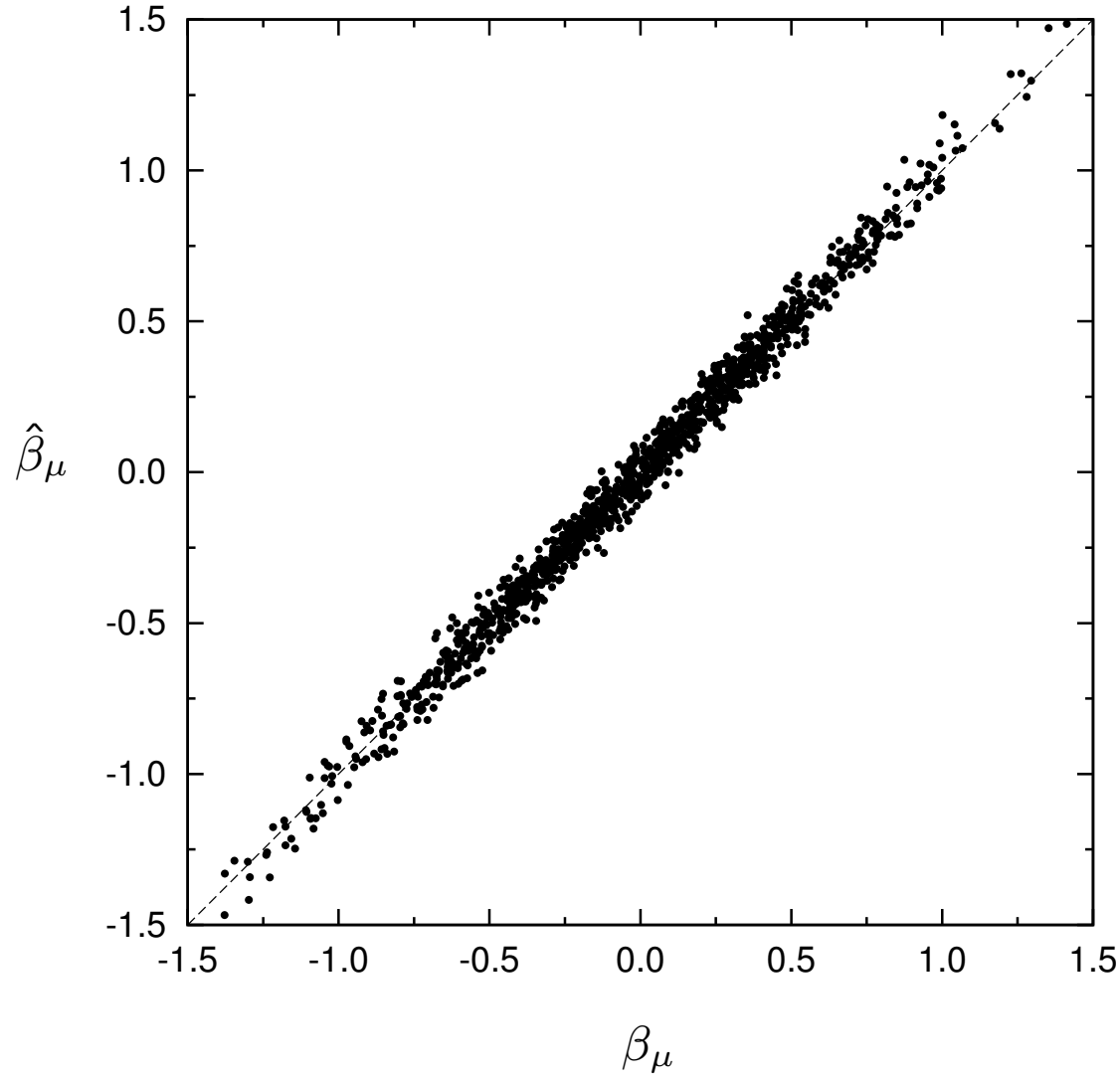
- put: $\hat{\boldsymbol{\beta}} = \boldsymbol{A}^{-\frac{1}{2}} \tilde{\boldsymbol{\beta}}$, $\hat{\lambda}(t) = \tilde{\lambda}(t) \mathrm{e}^{-\tilde{\boldsymbol{\beta}} \cdot \boldsymbol{A}^{-\frac{1}{2}} \bar{\boldsymbol{z}}}$

$$\tilde{\boldsymbol{\beta}} = \mathrm{argmax}_{\boldsymbol{\beta}} \sum_i \left\{ \boldsymbol{\beta} \cdot \tilde{\boldsymbol{z}}_i - \log \left[\sum_j \theta(t_j - t_i) \mathrm{e}^{\boldsymbol{\beta} \cdot \tilde{\boldsymbol{z}}_j} \right] \right\}, \qquad \tilde{\lambda}(t) = \frac{\sum_i \delta(t - t_i)}{\sum_i \theta(t_i - t) \mathrm{e}^{\tilde{\boldsymbol{\beta}} \cdot \tilde{\boldsymbol{z}}_i}}$$

(standard eqns for $\{\tilde{\boldsymbol{z}}_i\}$)

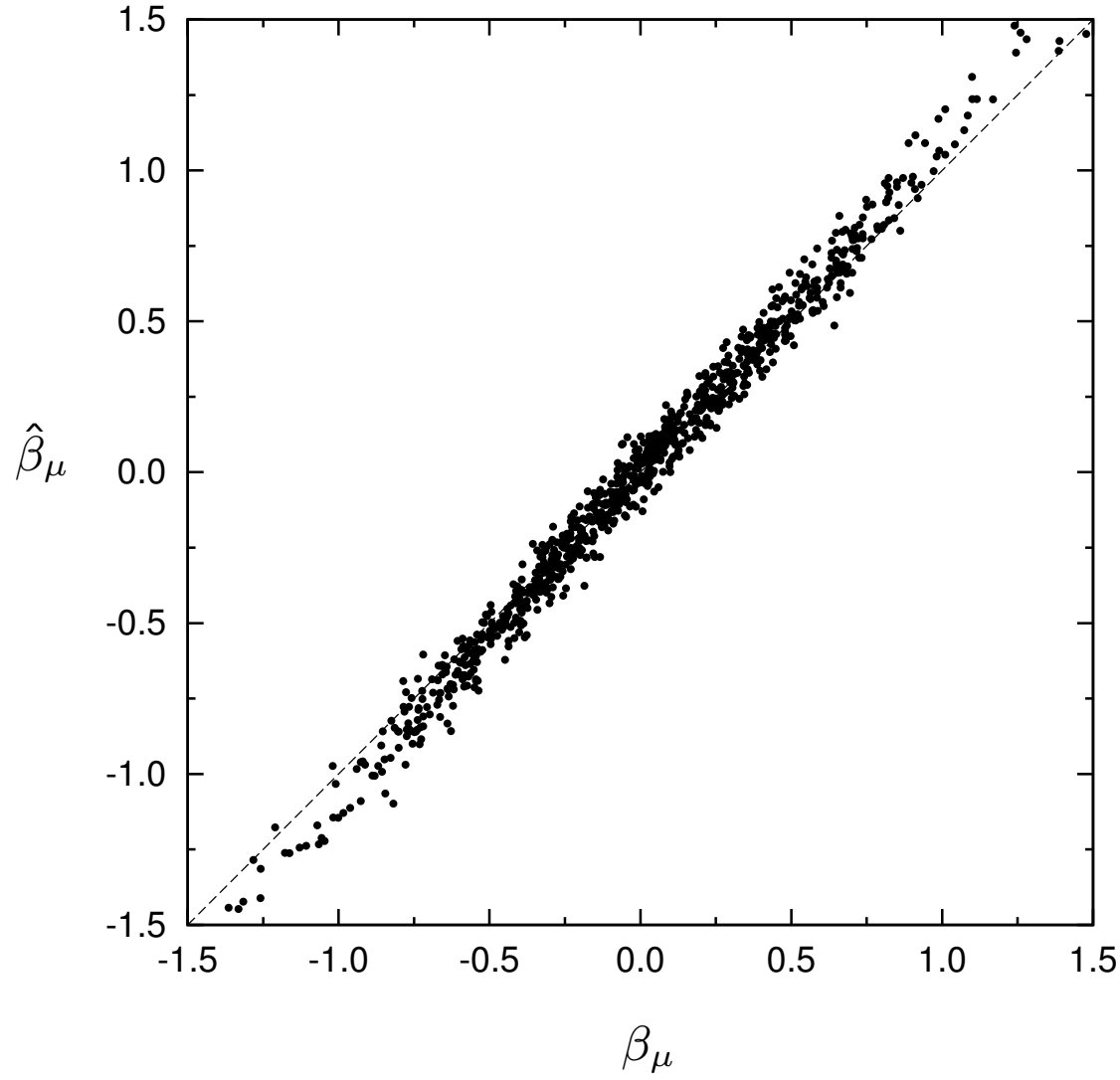
$N = 500$,
predicted versus true regression coefficients
(synthetic data, no censoring)

$p/N = 0.002$



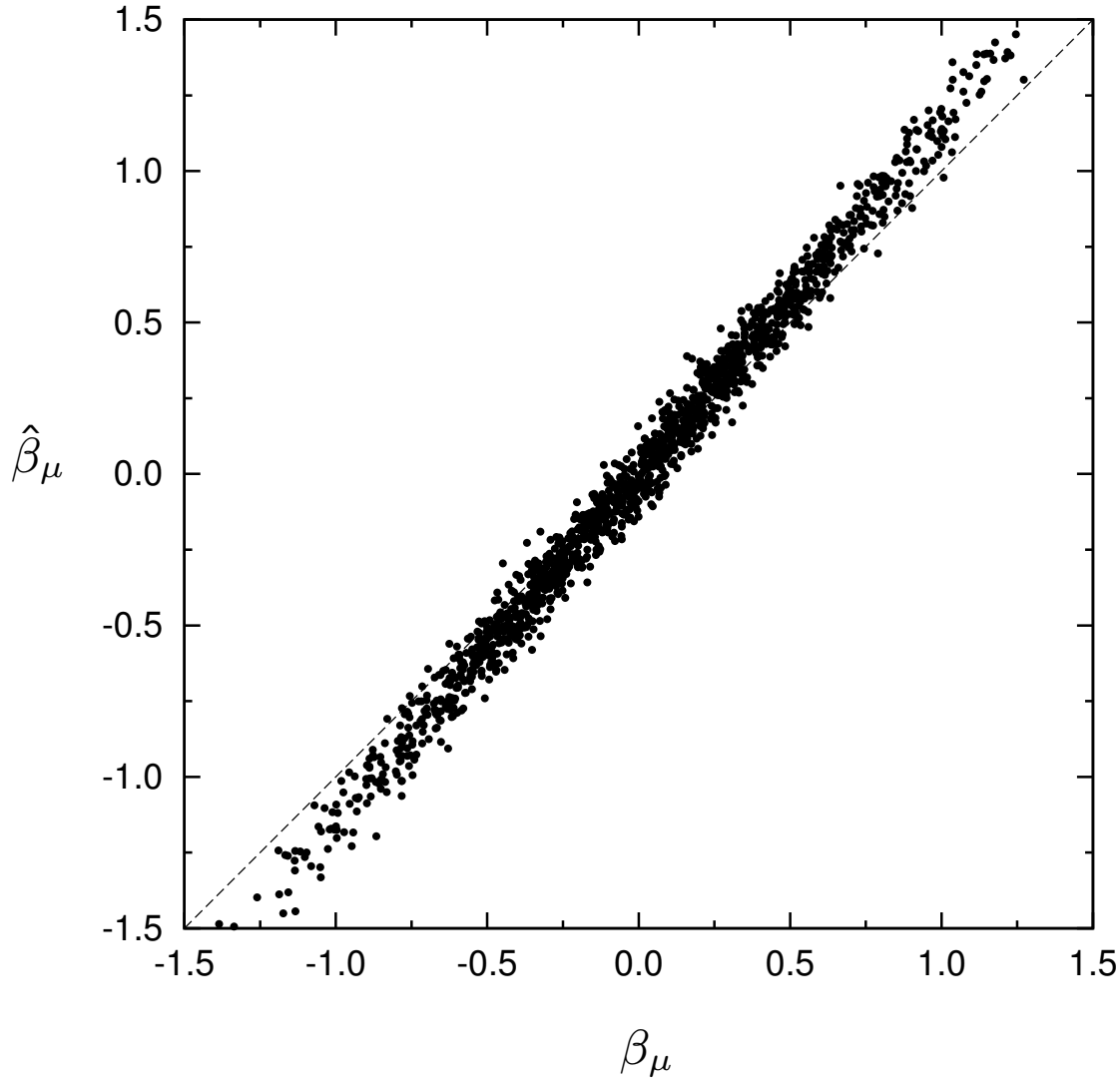
$N = 500$,
predicted versus true regression coefficients
(synthetic data, no censoring)

$p/N = 0.10$



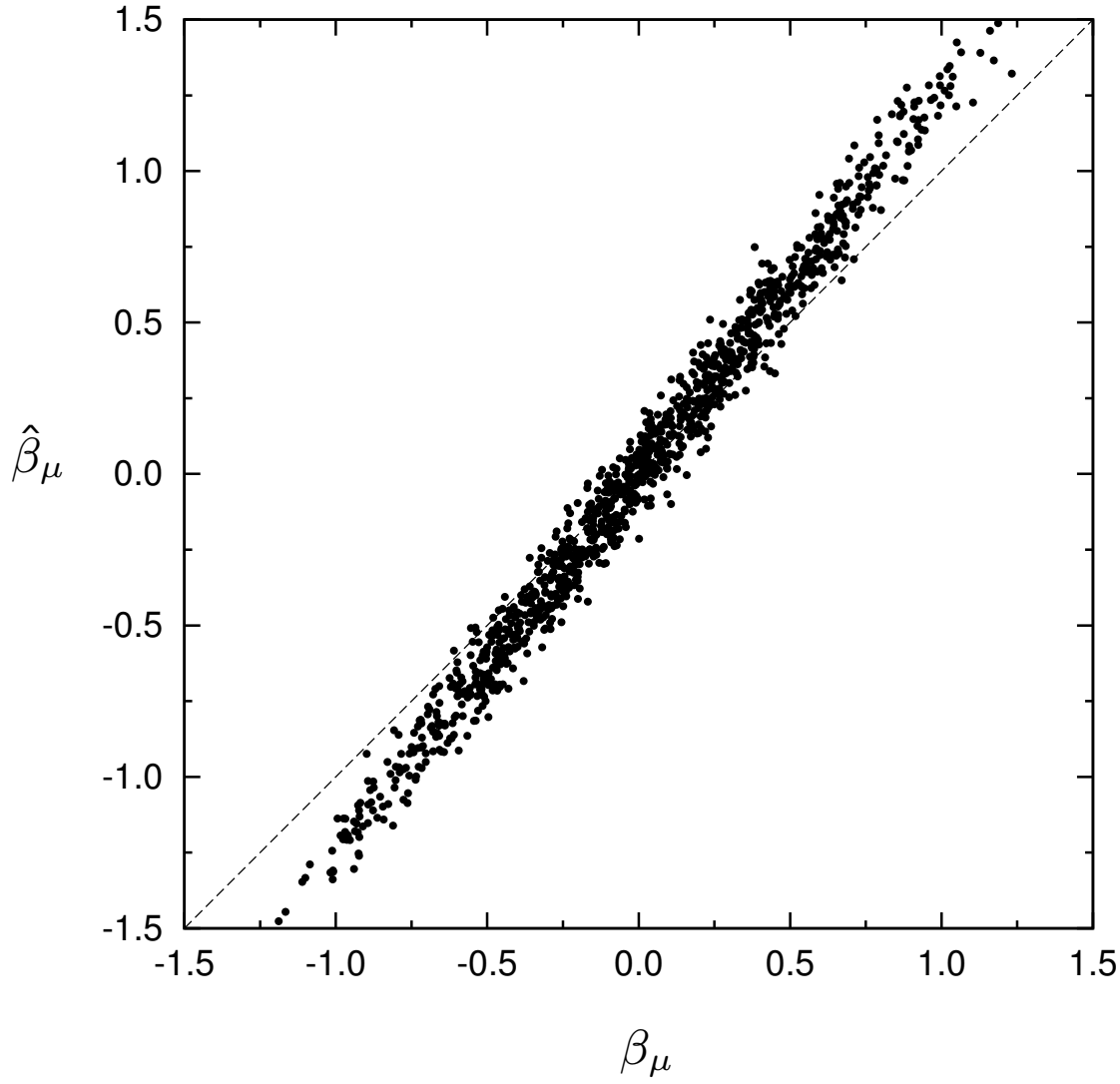
$N = 500$,
predicted versus true regression coefficients
(synthetic data, no censoring)

$p/N = 0.20$



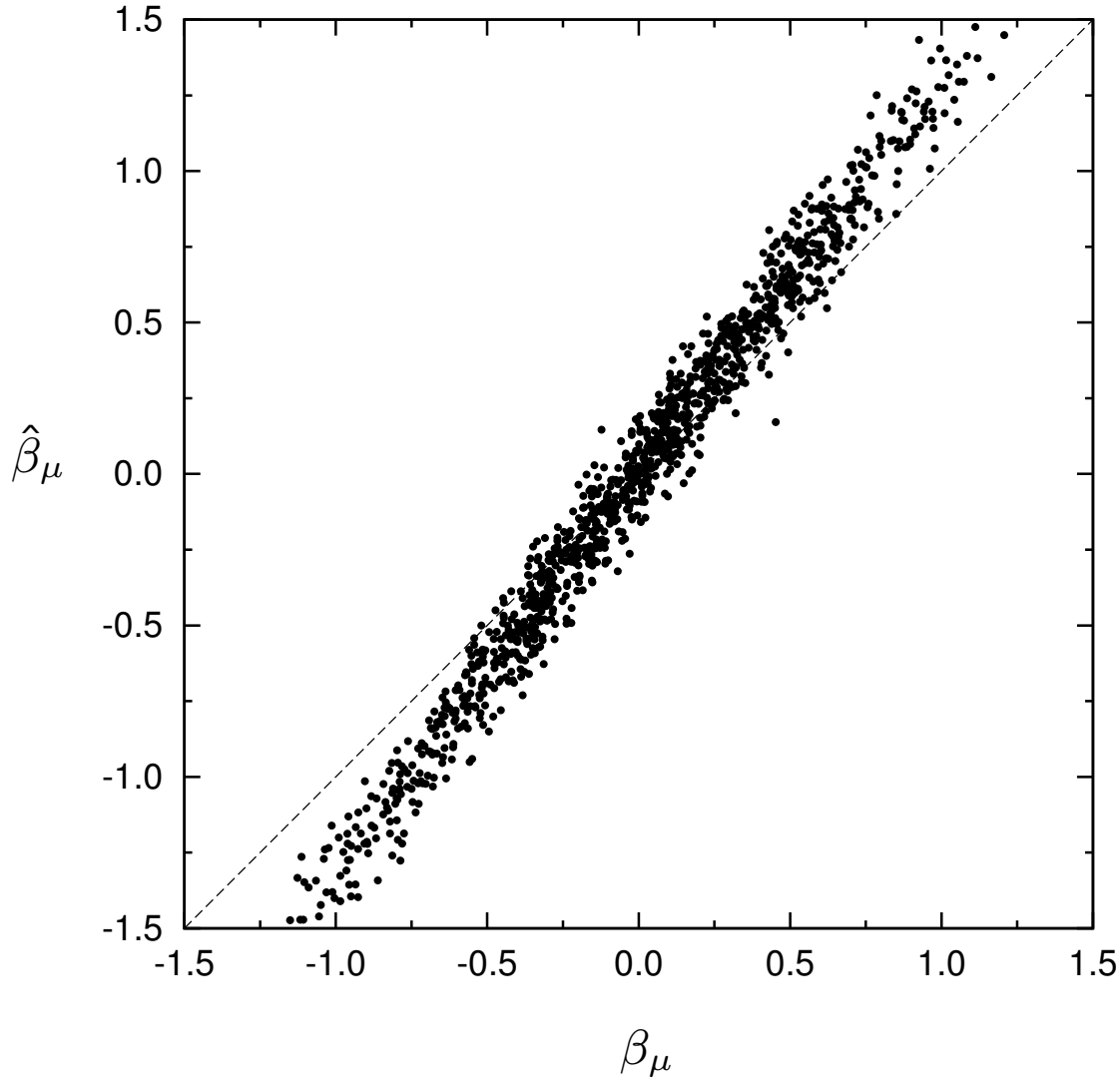
$N = 500$,
predicted versus true regression coefficients
(synthetic data, no censoring)

$p/N = 0.30$



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

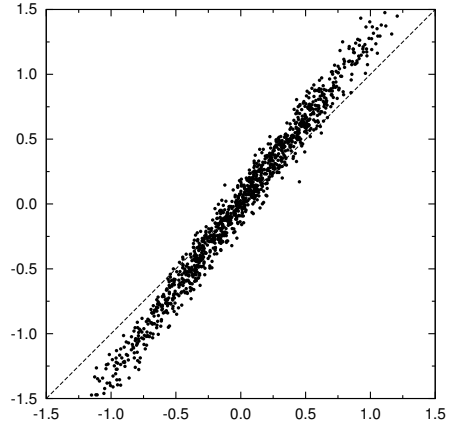
$p/N = 0.40$



Bad news

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

> *we always inflate associations*
> *(whether positive or negative)*

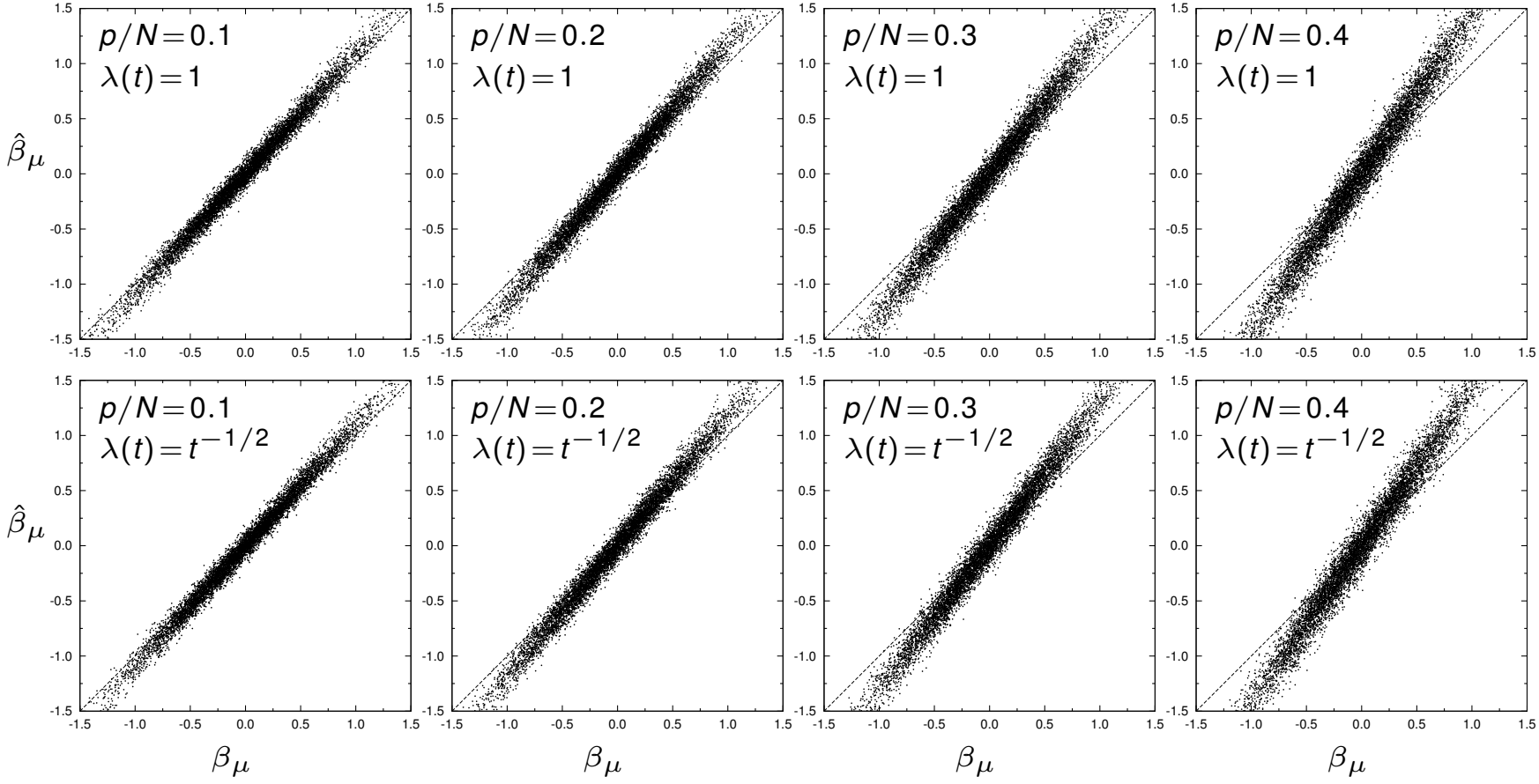


Good news

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

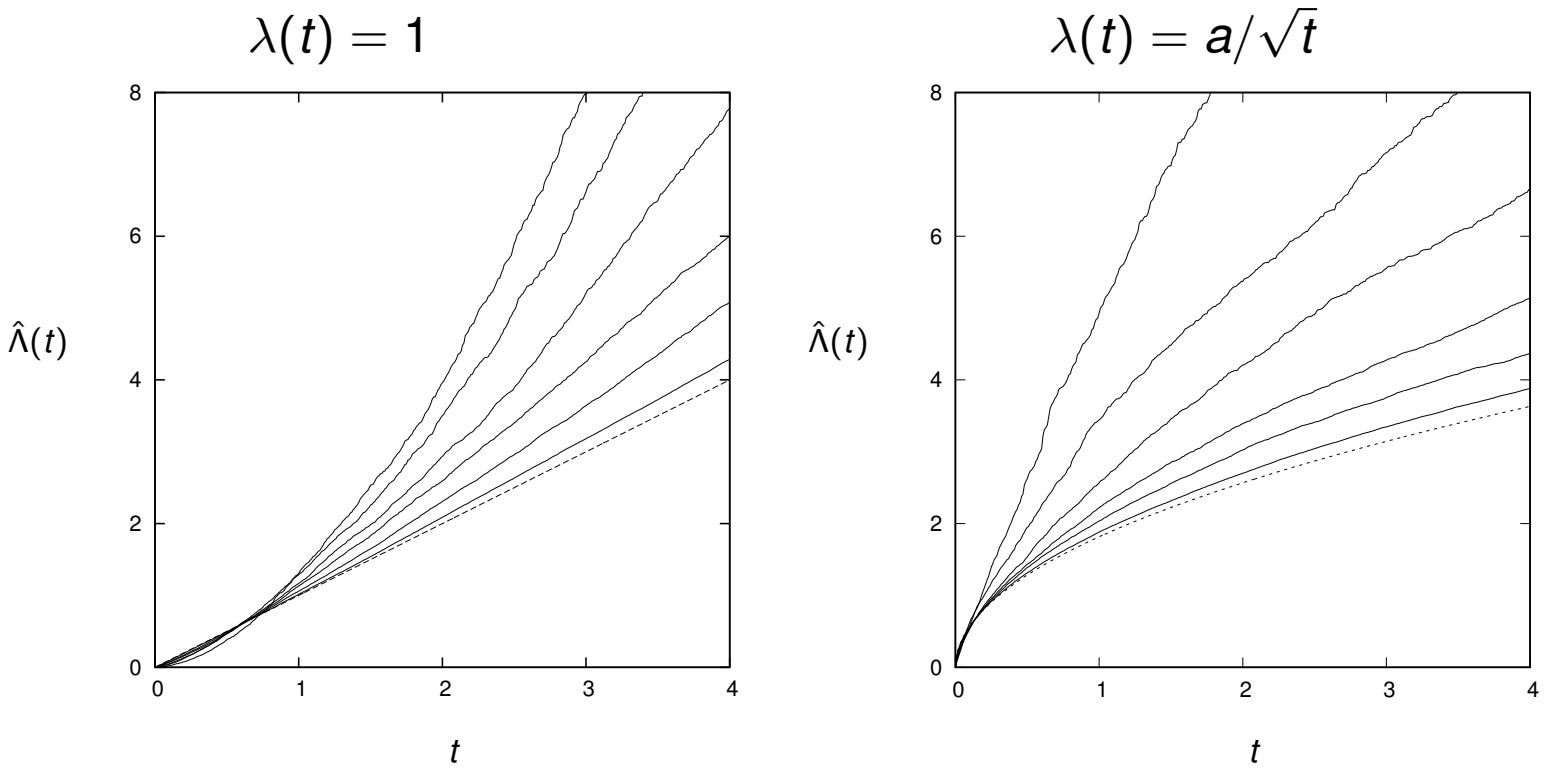
roadmap for research ...

- Predict impact of overfitting, in terms of
  - ratio p/N
  - covariate correlations
  - true parameters $\{\beta, \lambda(t)\}$
- 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 ...

- *ML regression*

  assumed model: $p_{\boldsymbol{\theta}} = \{p(t|\boldsymbol{z},\boldsymbol{\theta})\}$

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

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

  ML pushes $p(t|\boldsymbol{z},\boldsymbol{\theta})$ 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 $\boldsymbol{\theta}_{\mathrm{ML}} \neq \boldsymbol{\theta}^\star$ ✗

- *Barrier to overfitting theory*

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

Step 1 – what to compute?

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

  note:

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

  $$\text{if } \boldsymbol{\theta} = \boldsymbol{\theta}^\star: \quad D(\hat{p}_{\boldsymbol{\theta}^\star} || p_{\boldsymbol{\theta}}) = D(\hat{p}_{\boldsymbol{\theta}^\star} || p_{\boldsymbol{\theta}^\star}) \qquad \leftarrow \textit{only zero if } p \ll N$$

  Define:

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

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

- *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 β

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

- *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:

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

□

interpretation:
stochastic minimisation, with noise $\sim 1/\gamma$
(ground state = zero temp limit of free energy)

Step 3 – 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*

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

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

- *Application*

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

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

□

still completely general ...

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

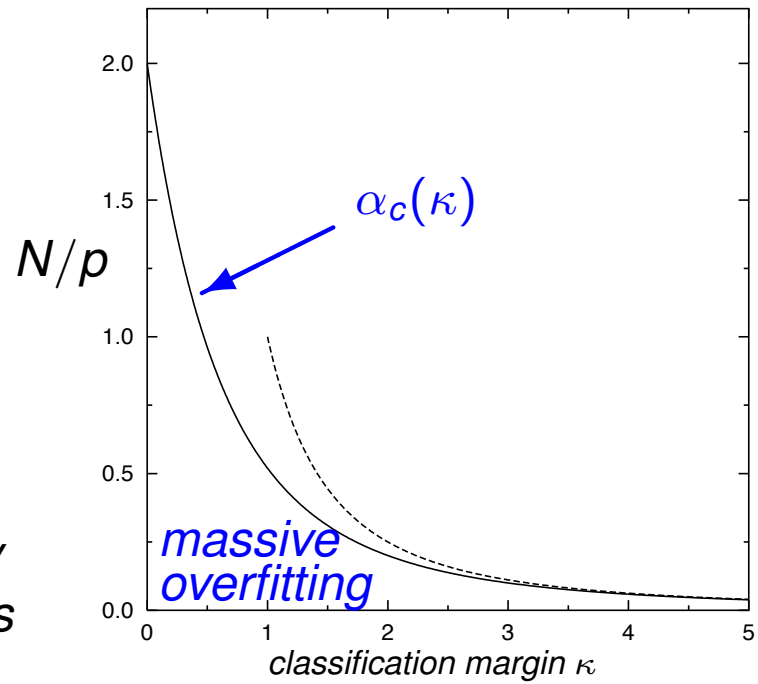
heterogeneous stochastic systems in physics,
biology, computer science, economics, ...

- *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

- Here $\boldsymbol{\theta} = \{\lambda, \boldsymbol{\beta}\}$, and

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

$$E_\gamma(\boldsymbol{\beta}^\star, \lambda^\star) = -\frac{\partial}{\partial\gamma} \lim_{n\to 0} \frac{1}{Nn} \log \int \{\mathrm{d}\lambda_1 \ldots \mathrm{d}\lambda_n\} \int \mathrm{d}\boldsymbol{\beta}^1 \ldots \mathrm{d}\boldsymbol{\beta}^n$$

$$\times \Big\{ \int \mathrm{d}\boldsymbol{z}\mathrm{d}t\ p(\boldsymbol{z}) p(t|\boldsymbol{z}, \lambda^\star, \boldsymbol{\beta}^\star) \prod_{\alpha=1}^{n} \Big[\frac{p(t|\boldsymbol{z}, \lambda_\alpha, \boldsymbol{\beta}^\alpha)}{p(t|\boldsymbol{z}, \lambda^\star, \boldsymbol{\beta}^\star)}\Big]^\gamma \Big\}^N$$

- short-hands:

$$p(t|\xi, \lambda) = \lambda(t)\mathrm{e}^{\xi - \exp(\xi)\int_0^t \mathrm{d}s\ \lambda(s)}, \qquad (\boldsymbol{\beta}^\star, \lambda^\star) = (\boldsymbol{\beta}^0, \lambda_0)$$

$$\boldsymbol{y} = (y_0, \ldots, y_p), \qquad p(\boldsymbol{y}|\boldsymbol{\beta}^0, \ldots, \boldsymbol{\beta}^n) = \int \mathrm{d}\boldsymbol{z}\ p(\boldsymbol{z}) \prod_{\alpha=0}^{n} \delta\Big[y_\alpha - \frac{\boldsymbol{\beta}^\alpha\cdot\boldsymbol{z}}{\sqrt{p}}\Big]$$

$$E_\gamma(\boldsymbol{\beta}^0, \lambda_0) = -\frac{\partial}{\partial\gamma} \lim_{n\to 0} \frac{1}{Nn} \log \int \{\mathrm{d}\lambda_1 \ldots \mathrm{d}\lambda_n\} \int \mathrm{d}\boldsymbol{\beta}^1 \ldots \mathrm{d}\boldsymbol{\beta}^n$$

$$\times \Big\{ \int \mathrm{d}\boldsymbol{y}\ p(\boldsymbol{y}|\boldsymbol{\beta}^0, \ldots, \boldsymbol{\beta}^n) \int \mathrm{d}t\ p(t|y_0, \lambda_0) \prod_{\alpha=1}^{n} \Big[\frac{p(t|y_\alpha, \lambda_\alpha)}{p(t|y_0, \lambda_0)}\Big]^\gamma \Big\}^N$$

- all $\{y_\alpha\}$: sums of zero-average Gaussian random vars, so

$$p(\boldsymbol{y}|\boldsymbol{\beta}^0,\ldots,\boldsymbol{\beta}^n) = \frac{\mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}[\{\boldsymbol{\beta}\}]\boldsymbol{y}}}{\sqrt{(2\pi)^{n+1}\mathrm{Det}\boldsymbol{C}[\{\boldsymbol{\beta}\}]}}$$

$$C_{\alpha\rho}[\{\boldsymbol{\beta}\}] = \frac{1}{p}\int \mathrm{d}\boldsymbol{z}\; p(\boldsymbol{z})(\boldsymbol{\beta}^\alpha\cdot\boldsymbol{z})(\boldsymbol{\beta}^\rho\cdot\boldsymbol{z}) = \frac{1}{p}\boldsymbol{\beta}^\alpha\cdot\boldsymbol{\beta}^\rho$$

- introduce for each (α,ρ):

$$1 = \int \mathrm{d}C_{\alpha\rho}\;\delta\Big[C_{\alpha\rho} - C_{\alpha\rho}[\{\boldsymbol{\beta}\}]\Big] = \int \frac{\mathrm{d}C_{\alpha\rho}\mathrm{d}\hat{C}_{\alpha\rho}}{2\pi/p}\;\mathrm{e}^{\mathrm{i}p\hat{C}_{\alpha\rho}\left[C_{\alpha\rho}-C_{\alpha\rho}[\{\boldsymbol{\beta}\}]\right]}$$

after simple manipulations,

$$E_\gamma(\boldsymbol{\beta}^0,\lambda_0) = -\frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{Nn}\log\int\{\mathrm{d}\lambda_1\ldots\mathrm{d}\lambda_n\}\int\frac{\mathrm{d}\boldsymbol{C}\mathrm{d}\hat{\boldsymbol{C}}\;\mathrm{e}^{\mathrm{i}p\sum_{\alpha\rho=0}^n\hat{C}_{\alpha\rho}C_{\alpha\rho}}}{(2\pi/p)^{(n+1)^2}}$$

$$\times\Big\{\int\frac{\mathrm{d}\boldsymbol{y}\;\mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}}}{\sqrt{(2\pi)^{n+1}\mathrm{Det}\boldsymbol{C}}}\int\mathrm{d}t\;p(t|y_0,\lambda_0)\prod_{\alpha=1}^n\Big[\frac{p(t|y_\alpha,\lambda_\alpha)}{p(t|y_0,\lambda_0)}\Big]^\gamma\Big\}^N$$

$$\times\int\mathrm{d}\boldsymbol{\beta}^1\ldots\mathrm{d}\boldsymbol{\beta}^n\;\mathrm{e}^{-\mathrm{i}\sum_{\alpha\rho=0}^n\hat{C}_{\alpha\rho}\boldsymbol{\beta}^\alpha\cdot\boldsymbol{\beta}^\rho}$$

Conversion to saddle-point problem

- define: $P(\beta_0) = \frac{1}{p}\sum_{\mu=1}^{p} \delta[\beta_0 - \beta_\mu^0]$, $S^2 = \int \mathrm{d}\beta_0 P(\beta_0)\beta_0^2$

$$E_\gamma(P,\lambda_0) = -\frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{Nn}\log\int\{\mathrm{d}\lambda_1\ldots\mathrm{d}\lambda_n\}\int\frac{\mathrm{d}\boldsymbol{C}\mathrm{d}\hat{\boldsymbol{C}}\;\mathrm{e}^{\mathrm{i}p\left(\sum_{\alpha\rho=0}^{n}\hat{C}_{\alpha\rho}C_{\alpha\rho}-\hat{C}_{00}S^2\right)}}{(2\pi/p)^{(n+1)^2}}$$
$$\times\Big\{\int\frac{\mathrm{d}\boldsymbol{y}\;\mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}}}{\sqrt{(2\pi)^{n+1}\mathrm{Det}\boldsymbol{C}}}\int\mathrm{d}t\;p(t|y_0,\lambda_0)\prod_{\alpha=1}^{n}\Big[\frac{p(t|y_\alpha,\lambda_\alpha)}{p(t|y_0,\lambda_0)}\Big]^\gamma\Big\}^N$$
$$\times\;\mathrm{e}^{p\int\mathrm{d}\beta_0\;P(\beta_0)\log\int\mathrm{d}\beta_1\ldots\mathrm{d}\beta_n\;\mathrm{e}^{-2\mathrm{i}\beta_0\sum_{\rho=1}^{n}\hat{C}_{0\rho}\beta_\rho-\mathrm{i}\sum_{\alpha\rho=1}^{n}\hat{C}_{\alpha\rho}\beta_\alpha\beta_\rho}}$$

- exchange order of $n\to 0$ and $N\to\infty$,
  with fixed $p/N=\zeta$: steepest descent integral ...

$$\lim_{N\to\infty}E_\gamma(P,\lambda_0) = \frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{n}\mathrm{extr}_{\boldsymbol{C},\hat{\boldsymbol{C}},\lambda_1,\ldots,\lambda_n}\Psi[\boldsymbol{C},\hat{\boldsymbol{C}};\lambda_1,\ldots,\lambda_n]$$
$$\Psi[\ldots] = -\mathrm{i}\zeta\Big[\sum_{\alpha\rho=0}^{n}\hat{C}_{\alpha\rho}C_{\alpha\rho}-\hat{C}_{00}S^2\Big]+\frac{1}{2}(n+1)\log(2\pi)+\frac{1}{2}\log\mathrm{Det}\boldsymbol{C}$$
$$-\zeta\int\mathrm{d}\beta_0\;P(\beta_0)\log\int\mathrm{d}\beta_1\ldots\mathrm{d}\beta_n\;\mathrm{e}^{-2\mathrm{i}\beta_0\sum_{\rho=1}^{n}\hat{C}_{0\rho}\beta_\rho-\mathrm{i}\sum_{\alpha\rho=1}^{n}\hat{C}_{\alpha\rho}\beta_\alpha\beta_\rho}$$
$$-\log\int\mathrm{d}\boldsymbol{y}\;\mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}}\int\mathrm{d}t\;p(t|y_0,\lambda_0)\prod_{\alpha=1}^{n}\Big[\frac{p(t|y_\alpha,\lambda_\alpha)}{p(t|y_0,\lambda_0)}\Big]^\gamma$$

- simple steps ...

$$C_{00} = S^2, \qquad \alpha, \rho = 1 \ldots n: \quad \hat{C}_{\alpha\rho} = -\frac{1}{2} \mathrm{i} D_{\alpha\rho}, \quad \hat{C}_{0\rho} = -\frac{1}{2} \mathrm{i} d_\rho$$

with $D_{\alpha\rho}, d_\rho \in \mathbb{R}$, and $\boldsymbol{D} = \{D_{\alpha\rho}\}$ positive definite,

$$\begin{aligned}
\Psi[\ldots] = & -\frac{1}{2}\zeta \sum_{\alpha\rho=1}^{n} D_{\alpha\rho} C_{\alpha\rho} - \zeta \sum_{\rho=1}^{n} d_\rho C_{0\rho} - \frac{1}{2}\zeta S^2 \sum_{\alpha\rho=1}^{n} d_\alpha (\boldsymbol{D}^{-1})_{\alpha\rho} d_\rho \\
& + \frac{1}{2}(n+1)\log(2\pi) + \frac{1}{2}\log \mathrm{Det}\boldsymbol{C} - \zeta \log \int \mathrm{d}\beta_1 \ldots \mathrm{d}\beta_n \, \mathrm{e}^{-\frac{1}{2}\sum_{\alpha\rho=1}^{n} D_{\alpha\rho}\beta_\alpha\beta_\rho} \\
& - \log \int \mathrm{d}\boldsymbol{y} \, \mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}} \int \mathrm{d}t \; p(t|y_0,\lambda_0) \prod_{\alpha=1}^{n} \left[\frac{p(t|y_\alpha,\lambda_\alpha)}{p(t|y_0,\lambda_0)}\right]^\gamma
\end{aligned}$$

- variation with respect to $\{d_\alpha\}$: $d_\alpha = -\sum_\rho D_{\alpha\rho} C_{0\rho}/S^2$

$$\begin{aligned}
\Psi[\ldots] \quad = \quad & -\frac{1}{2}\zeta \sum_{\alpha\rho=1}^{n} D_{\alpha\rho}\left[C_{\alpha\rho} - \frac{C_{0\alpha}C_{0\rho}}{S^2}\right] + \frac{1}{2}(n+1)\log(2\pi) + \frac{1}{2}\log \mathrm{Det}\boldsymbol{C} \\
& - \log \int \mathrm{d}\boldsymbol{y} \, \mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}} \int \mathrm{d}t \; p(t|y_0,\lambda_0) \prod_{\alpha=1}^{n} \left[\frac{p(t|y_\alpha,\lambda_\alpha)}{p(t|y_0,\lambda_0)}\right]^\gamma \\
& - \zeta \log \int \mathrm{d}\beta_1 \ldots \mathrm{d}\beta_n \, \mathrm{e}^{-\frac{1}{2}\sum_{\alpha\rho=1}^{n} D_{\alpha\rho}\beta_\alpha\beta_\rho}
\end{aligned}$$

- eqns depend on $P(\beta_0)$ only via $S^2 = \int \mathrm{d}\beta_0 \; P(\beta_0)\beta_0^2$
- variation with respect to $\boldsymbol{D}$: $(\boldsymbol{D}^{-1})_{\alpha\rho} = C_{\alpha\rho} - C_{0\alpha}C_{0\rho}/S^2$,

$$E_\gamma(S,\lambda_0) = \frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{n}\mathrm{extr}_{\boldsymbol{C};\lambda_1,\ldots,\lambda_n}\Psi[\boldsymbol{C};\lambda_1,\ldots,\lambda_n]$$

$$\Psi[\boldsymbol{C};\lambda_1,\ldots,\lambda_n] = \frac{1}{2}\log\mathrm{Det}\boldsymbol{C} - \frac{1}{2}\zeta\log\mathrm{Det}\boldsymbol{C}'$$
$$- \log\int\frac{\mathrm{d}\boldsymbol{y}}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}}\int\mathrm{d}t\; p(t|y_0,\lambda_0)\prod_{\alpha=1}^{n}\Big[\frac{p(t|y_\alpha,\lambda_\alpha)}{p(t|y_0,\lambda_0)}\Big]^\gamma$$

notes:

modulo terms that vanish due to $n\to 0$ or $\partial/\partial\gamma$,

extremisation over $\boldsymbol{C}$: subject to $C_{00} = S^2$,

$\boldsymbol{C}'$: $n\times n$ matrix, $C'_{\alpha\rho} = C_{\alpha\rho} - C_{0\alpha}C_{0\rho}/S^2$ $(\alpha,\rho = 1\ldots n)$

to proceed:
ansatz for form of $\{C_{\alpha\beta},\lambda_\alpha\}$

Replica symmetric saddle points

solution space of the regression algorithm *ergodic*,
set of equivalent minima in regression parameter space *connected*

$$\forall \alpha, \rho = 1 \ldots n: \qquad \lambda_\alpha(t) = \lambda(t), \quad C_{0\alpha} = c_0,$$
$$C_{\alpha\rho} = C\delta_{\alpha\rho} + c(1-\delta_{\alpha\rho})$$

$$E_\gamma(S, \lambda_0) = \frac{\partial}{\partial\gamma} \mathrm{extr}_{C,c,c_0;\lambda} \Psi_{\mathrm{RS}}[C, c, c_0; \lambda]$$

$$\Psi_{\mathrm{RS}}[C, c, c_0; \lambda] = \lim_{n\to 0} \frac{1}{n} \Big\{ \frac{1}{2} \log \mathrm{Det} \boldsymbol{C} - \frac{1}{2}\zeta \log \mathrm{Det} \boldsymbol{C}'$$
$$- \log \int \frac{\mathrm{d}\boldsymbol{y}}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}} \int \mathrm{d}t\; p(t|y_0, \lambda_0) \prod_{\alpha=1}^{n} \Big[\frac{p(t|y_\alpha, \lambda)}{p(t|y_0, \lambda_0)} \Big]^\gamma \Big\}$$

$$\boldsymbol{C} = \begin{pmatrix} S^2 & c_0 & \cdots & \cdots & c_0 \\ c_0 & C & c & \cdots & c \\ \vdots & c & C & \cdots & c \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_0 & c & \cdots & c & C \end{pmatrix}$$

- eigenvectors and eigenvalues of $\boldsymbol{C}$:

$$(u, v, \ldots, v): \quad \mu_\pm = \frac{1}{2}\Big\{C + (n-1)c + S^2 \pm \sqrt{[C+(n-1)c - S^2]^2 + 4nc_0^2}\Big\}$$

$$(0, v_1, \ldots, v_n): \quad \sum_{\alpha=1}^{n} v_\alpha = 0, \quad \mu = C - c \quad (\text{multiplicity } n-1)$$

so

$$\begin{aligned} \log \mathrm{Det}\boldsymbol{C} &= \log[(C-c)^{n-1}\mu_+\mu_-] \\ &= \log S^2 + n\log(C-c) + n\frac{c - c_0^2/S^2}{C-c} + \mathcal{O}(n^2) \end{aligned}$$

- form of $\boldsymbol{C}^{-1}$:

$$\boldsymbol{C}^{-1} = \begin{pmatrix} d_{00} & d_0 & \cdots & \cdots & d_0 \\ d_0 & D & d & \cdots & d \\ \vdots & d & D & \cdots & d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_0 & d & \cdots & d & D \end{pmatrix}$$

$$d_{00} = \frac{C + (n-1)c}{S^2(C + (n-1)c) - nc_0^2}, \qquad d_0 = -\frac{c_0}{S^2(C + (n-1)c) - nc_0^2}$$

$$d = \frac{1}{C-c}\,\frac{c_0^2 - cS^2}{S^2(C + (n-1)c) - nc_0^2}, \qquad D = d + \frac{1}{C-c}$$

- eigenvectors and eigenvalues of $\boldsymbol{C}'$:

$$\begin{aligned} (1,\ldots,1): &\qquad \textit{eigenvalue}\ \ C-c-nc_0^2/S^2+nc \\ (1,\ldots,1)^\perp: &\qquad \textit{eigenvalue}\ \ C-c\ \ \ (n-1\ \textit{fold degenerate}) \end{aligned}$$

so

$$\begin{aligned} \log \mathrm{Det}\boldsymbol{C}' &= (n-1)\log(C-c)+\log\Big(C-c+n(c-c_0^2/S^2)\Big) \\ &= n\Big[\log(C-c)+(c-c_0^2/S^2)(C-c)\Big]+\mathcal{O}(n^2) \end{aligned}$$

- insert into $\Psi_{\rm RS}$,
  expand for small n:

$$\Psi_{\rm RS}[C,c,c_0;\lambda]=\frac{1}{2}(1-\zeta)\Big[\log(C-c)+\frac{c-c_0^2/S^2}{C-c}\Big]+\frac{1}{2}\frac{c_0^2/S^2}{C-c}-\frac{1}{2}\log(C-c)$$
$$-\frac{1}{2}\log(2\pi)-\int \mathrm{D}z\mathrm{D}y_0\int \mathrm{d}t\ p(t|Sy_0,\lambda_0)$$
$$\times\log\int \mathrm{D}y\ \mathrm{e}^{y[y_0c_0/S\sqrt{C-c}+z\sqrt{(c-c_0^2/S^2)/(C-c)}]}\Big(\frac{p(t|y\sqrt{C-c},\lambda)}{p(t|Sy_0,\lambda_0)}\Big)^\gamma$$

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

- transform RS order pars: $u = \sqrt{C-c},\ v = \sqrt{c-c_0^2/S^2},\ w = c_0/S,$
  transform integration var: $y \to y + (wy_0 + vz)/u,$

$$E_\gamma(S,\lambda_0) = \int \mathrm{D}y_0 \int \mathrm{d}t\ p(t|Sy_0,\lambda_0) \log p(t|Sy_0,\lambda_0)$$

$$-\frac{\partial}{\partial\gamma}\mathrm{extr}_{u,v,w;\lambda}\Big\{\zeta\Big(\frac{v^2}{2u^2} + \log u\Big)$$

$$+ \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\ p(t|Sy_0,\lambda_0) \log \int \mathrm{D}y\ p^\gamma(t|uy + wy_0 + vz,\lambda)\Big\}$$

$$= \int \mathrm{D}y_0 \int \mathrm{d}t\ p(t|Sy_0,\lambda_0)\Big\{\log p(t|Sy_0,\lambda_0)$$

$$- \int \mathrm{D}z\left[\frac{\int \mathrm{D}y\ p^\gamma(t|uy + wy_0 + vz,\lambda) \log p(t|uy + wy_0 + vz,\lambda)}{\int \mathrm{D}y\ p^\gamma(t|uy + wy_0 + vz,\lambda)}\right]\Big\}$$

$\{u, v, w; \lambda\}$: evaluated at saddle point of

$$\Psi_{\mathrm{RS}}(u,v,w;\lambda) = \zeta\Big(\frac{v^2}{2u^2} + \log u\Big)$$

$$+ \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\ p(t|Sy_0,\lambda_0) \log \int \mathrm{D}y\ p^\gamma(t|uy + wy_0 + vz,\lambda)$$

Physical interpretation of order parameters

- usual replica manipulations:
  (fraction form, or weak perturbations):

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

  transformed order pars:

$$u^2 = \lim_{p\to\infty} \frac{1}{p}\langle\langle\boldsymbol{\beta}^2\rangle - \langle\boldsymbol{\beta}\rangle^2\rangle_{\mathcal{D}}, \qquad w = \lim_{p\to\infty} \frac{1}{\sqrt{p}} \frac{\boldsymbol{\beta}^\star \cdot \langle\langle\boldsymbol{\beta}\rangle\rangle_{\mathcal{D}}}{|\boldsymbol{\beta}^\star|}$$

$$v^2 = \lim_{p\to\infty} \frac{1}{p}\left[\langle\langle\boldsymbol{\beta}\rangle^2\rangle_{\mathcal{D}} - \left(\frac{\boldsymbol{\beta}^\star \cdot \langle\langle\boldsymbol{\beta}\rangle\rangle_{\mathcal{D}}}{|\boldsymbol{\beta}^\star|}\right)^2\right]$$

- unique ML point: $\lim_{\gamma\to\infty} u = 0$
  (true for Cox regression, minimization of convex function)

  perfect regression:
  $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$ for all $\mathcal{D}, \boldsymbol{\beta}^\star$:

$$c_0 = c = C = S^2, \qquad u = v = 0,\ w = S$$

- if $\langle\boldsymbol{\beta}\rangle \approx \kappa\boldsymbol{\beta}^\star + \boldsymbol{\xi}$,
  with $\langle\boldsymbol{\xi}\rangle_{\mathcal{D}} = \mathbf{0}$, $\langle\boldsymbol{\xi}^2\rangle_{\mathcal{D}}/p = \sigma^2$:

$$\kappa = w/S, \qquad \sigma = v \quad (\textit{slope and width of parameter cloud})$$

RS saddle point eqns

- eqns for (u, v, w), using $\partial \log p(t|\xi)/\partial\xi = 1 - \mathrm{e}^{\xi}\Lambda(t)$:

$$\frac{\zeta}{\gamma u^2}\left(\frac{v^2}{\gamma u^2} - \frac{1}{\gamma}\right) = \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\; p(t|Sy_0, \lambda_0) \times \frac{\int \mathrm{D}y\; p^{\gamma}(t|uy + wy_0 + vz, \lambda)\left[[1 - \mathrm{e}^{uy+wy_0+vz}\Lambda(t)]^2 - \gamma^{-1}\mathrm{e}^{uy+wy_0+vz}\Lambda(t)\right]}{\int \mathrm{D}y\; p^{\gamma}(t|uy + wy_0 + vz, \lambda)}$$

$$\frac{\zeta v}{\gamma u^2} = \int \mathrm{D}z\mathrm{D}y_0\; z \int \mathrm{d}t\; p(t|Sy_0, \lambda_0)\Lambda(t)\frac{\int \mathrm{D}y\; p^{\gamma}(t|uy + wy_0 + vz, \lambda)\mathrm{e}^{uy+wy_0+vz}}{\int \mathrm{D}y\; p^{\gamma}(t|uy + wy_0 + vz, \lambda)}$$

$$0 = \int \mathrm{D}z\mathrm{D}y_0\; y_0 \int \mathrm{d}t\; p(t|Sy_0, \lambda_0)\Lambda(t)\frac{\int \mathrm{D}y\; p^{\gamma}(t|uy + wy_0 + vz, \lambda)\mathrm{e}^{uy+wy_0+vz}}{\int \mathrm{D}y\; p^{\gamma}(t|uy + wy_0 + vz, \lambda)}$$

- functional order parameter eqn, $\delta\Psi_{\mathrm{RS}}(u, v, w; \lambda)/\delta\lambda(s) = 0$:

$$\frac{p(s)}{\lambda(s)} = \int \mathrm{D}z\mathrm{D}y_0 \int_s^{\infty} \mathrm{d}t\; p(t|Sy_0, \lambda_0)\frac{\int \mathrm{D}y\; p^{\gamma}(t|uy + wy_0 + vz, \lambda)\mathrm{e}^{uy+wy_0+vz}}{\int \mathrm{D}y\; p^{\gamma}(t|uy + wy_0 + vz, \lambda)}$$

- correct scaling: $u = \tilde{u}/\sqrt{\gamma}$,
  tricky manipulations to get $\gamma \to \infty$ limit:

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

$W(z)$: Lambert's W-function
(inverse of $f(x) = x\mathrm{e}^x$)

- overfitting measure:

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

- $\zeta \to 0$: correct soln, no overfitting $(\tilde{u}=v=0,\ w=S)$
  $\zeta \to 1$: phase transition $(\tilde{u}, v, w \to \infty)$

preparations:

- write functional eqn in form involving $\Lambda(t)$ only:

$$\Lambda(t) = \int_0^t \mathrm{d}t'\; p(t')\Big\{ \int \mathrm{D}z\mathrm{D}y_0 \int_{t'}^{\infty} \mathrm{d}t''\; p(t''|Sy_0,\lambda_0)\, \frac{W\big(\tilde{u}^2 \mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t'')\big)}{\tilde{u}^2\Lambda(t'')} \Big\}^{-1}$$

- transform integration over t into
  integration over survival function $s = \exp[-\mathrm{e}^{Sy_0}\Lambda_0(t)]$,

  short-hand $L(t) = \tilde{u}^2\mathrm{e}^{\tilde{u}^2}\Lambda(t)$,

$$\zeta v^2 = \int \mathrm{D}y_0\mathrm{D}z \int_0^1 \mathrm{d}s\, \Big[\tilde{u}^2 - W\Big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\Big)\Big]^2$$

$$\zeta = \int \mathrm{D}y_0\mathrm{D}z \int_0^1 \mathrm{d}s\, \left\{ \frac{W\Big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\Big)}{1+W\Big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\Big)} \right\}$$

$$\frac{\zeta w}{S} = -\int \mathrm{D}y_0\mathrm{D}z \int_0^1 \mathrm{d}s\, \big[1+\log(s)\big]\, W\Big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\Big)$$

$$L(t) = \tilde{u}^2 \int_0^t \mathrm{d}t'\; p(t') \left\{ \int \mathrm{D}y_0\mathrm{D}z \int_0^1 \mathrm{d}s'\; \frac{\theta[t(s',y_0)-t']}{L(t(s',y_0))}\; W\Big(\mathrm{e}^{wy_0+vz}L(t(s',y_0))\Big) \right\}^{-1}$$

- main mathematical challenge: functional eqn

  write $L(t) = \Phi(\Lambda_0(t))$:

$$\frac{\tilde{u}^2 g(x)}{\mathrm{d}\log\Phi(x)/\mathrm{d}x}\,\frac{\mathrm{d}}{\mathrm{d}x}\log\left(\frac{\mathrm{d}\Phi(x)/\mathrm{d}x}{g(x)}\right) = \int \mathrm{D}y_0\ \mathrm{e}^{Sy_0 - x\exp(Sy_0)} \int \mathrm{D}z\ W\left(\mathrm{e}^{wy_0+vz}\Phi(x)\right)$$

$$g(x) = \int \mathrm{D}y\ \mathrm{e}^{Sy - x\exp(Sy)}$$

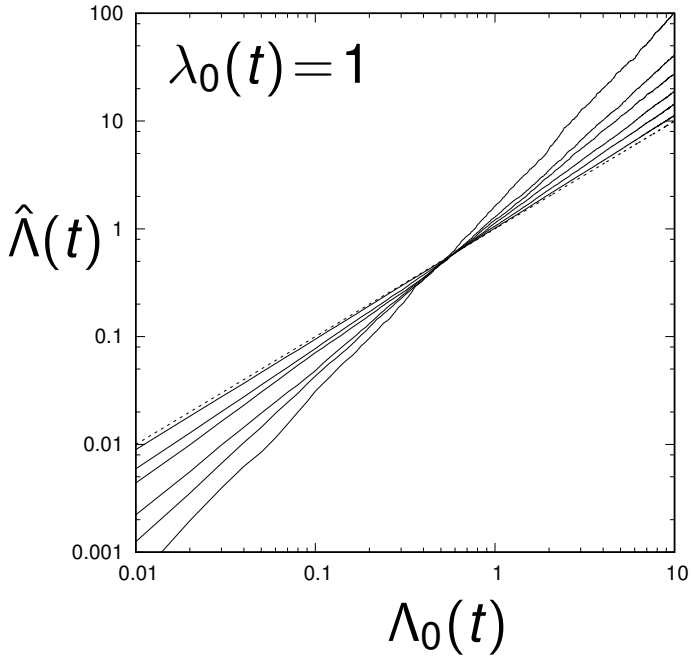
- asymptotic form, $x \to \infty$, using

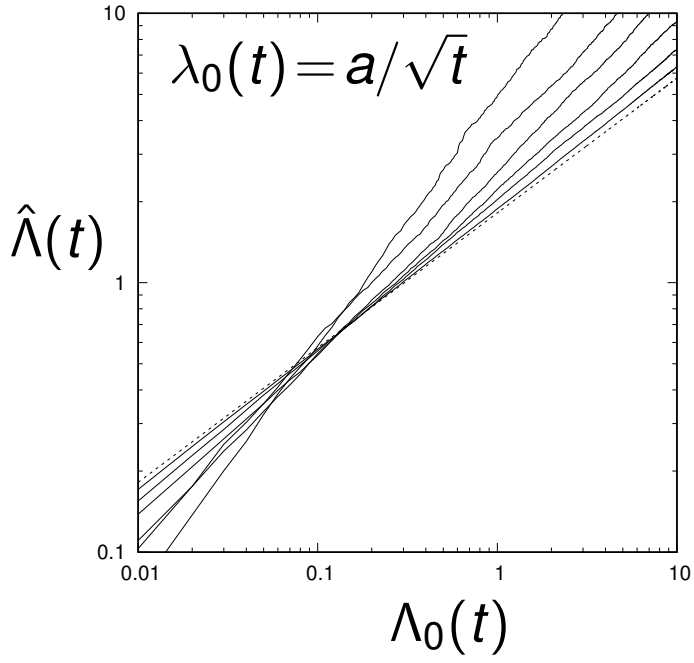
$$W(z) = \log z - \log(\log z) + \mathcal{O}(\log(\log z)/\log z) \qquad (z \to \infty)$$
$$\log g(x) = -\tfrac{1}{2S^2}(\log x)^2 + \tfrac{1}{S^2}\log x.\log(\log x) + \mathcal{O}(\log x) \qquad (x \to \infty)$$

$$\Phi(x) = \rho\log x + (1-\rho)\log\log x + o(\log\log x)$$
$$\rho = \frac{w}{2S}\left(1+\sqrt{1+4\tilde{u}^2/w^2}\right)$$

  hence

$$t \gg 1: \qquad \log\Lambda(t) = \rho\log\Lambda_0(t) + (1-\rho)\log(\log\Lambda_0(t)) + \ldots$$
$$\textit{leading order}: \quad \Lambda(t) \approx \Lambda_0^{\rho}(t)$$





simulation results:

$\zeta = 0.05, 0.15, 0.25, 0.35, 0.45, 0.55$

confirms $\log \Lambda(t) \approx \rho \log \Lambda_0(t)$,

with $\rho = \rho(\zeta)$

Variational approximation

$\Lambda(t) = k\Lambda_0^\rho(t)$,
insert into $\Psi_{\rm RS}$, extremize over $(k, \rho, \tilde{u}, v, w)$

write $q = k\tilde{u}^2 \mathrm{e}^{\tilde{u}^2}$,
further manipulations of integrals ...

$$\zeta v^2 = \int \mathrm{D}x \int_0^1 \mathrm{d}s \left[\tilde{u}^2 - W\left(q\mathrm{e}^{x\sigma(v,w,\rho)} \log^\rho(1/s)\right)\right]^2$$

$$\zeta = \int \mathrm{D}x \int_0^1 \mathrm{d}s \, \frac{W\left(q\mathrm{e}^{x\sigma(v,w,\rho)} \log^\rho(1/s)\right)}{1+W\left(q\mathrm{e}^{x\sigma(v,w,\rho)} \log^\rho(1/s)\right)}$$

$$\frac{\zeta w}{S} = -\int \mathrm{D}x \int_0^1 \mathrm{d}s \, \left[1+\log(s)\right] \, W\left(q\mathrm{e}^{x\sigma(v,w,\rho)} \log^\rho(1/s)\right)$$

$$\tilde{u}^2 = \int \mathrm{D}x \int_0^1 \mathrm{d}s \, W\left(q\mathrm{e}^{x\sigma(v,w,\rho)} \log^\rho(1/s)\right)$$

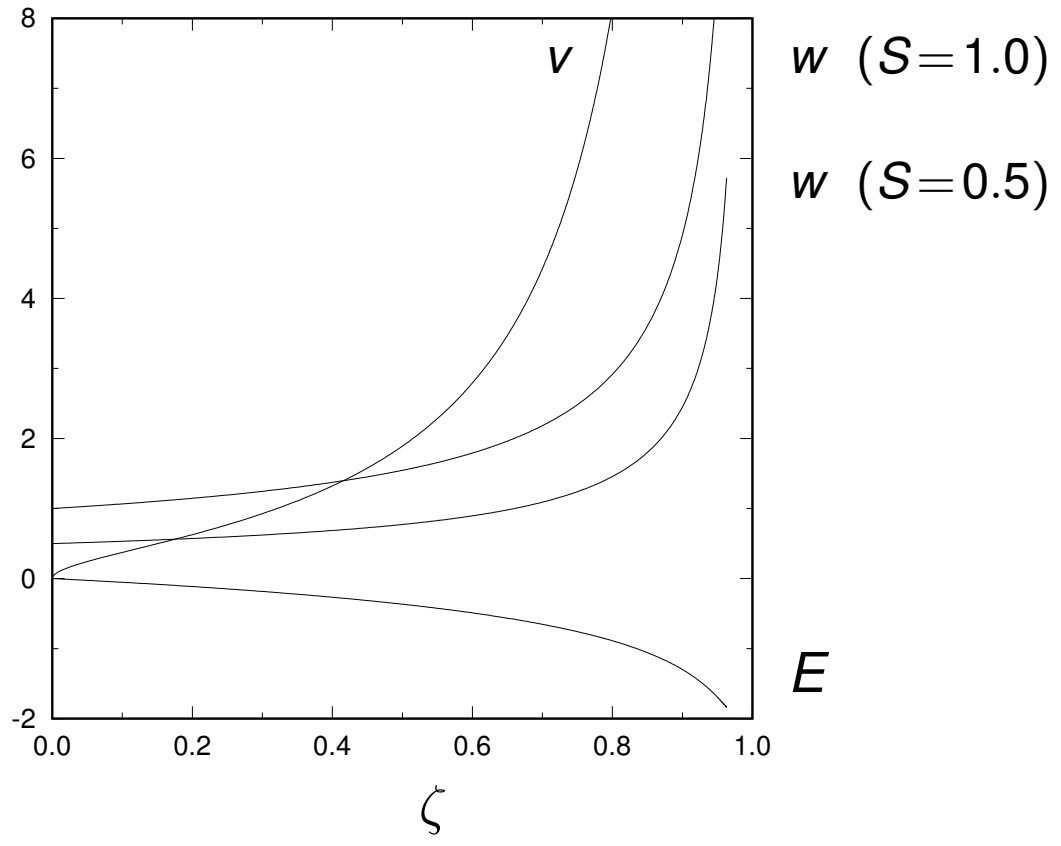
$$\frac{\tilde{u}^2}{\rho} = \int \mathrm{D}x \int_0^1 \mathrm{d}s \, W\left(q\mathrm{e}^{x\sigma(v,w,\rho)} \log^\rho(1/s)\right) \log\log(1/s) - S(w-\rho S)\zeta + \tilde{u}^2 C_{\rm E}$$

$C_{\rm E}$: Euler's constant

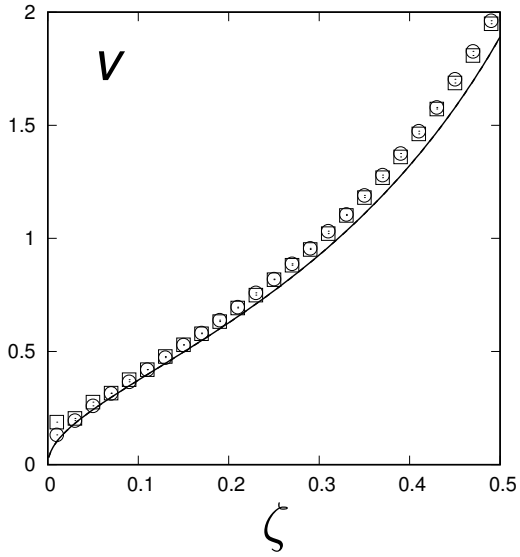
- similarly:

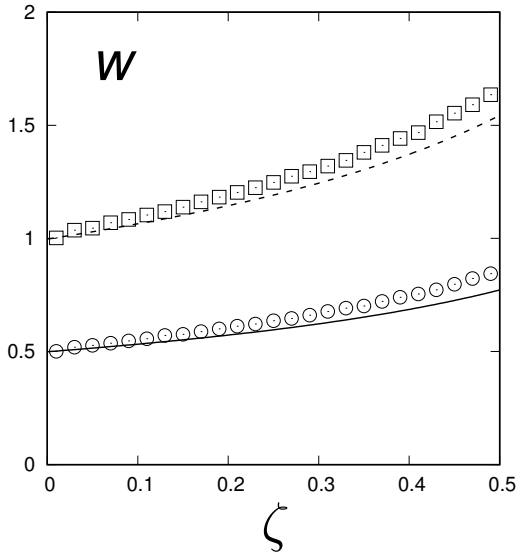
$$\begin{aligned} E(S,\lambda_0) &= \int \mathrm{d}t\; p(t) \log\left[\lambda_0(t)/\lambda(t)\right] \\ &= -\log k - \log\rho + (\rho-1)C_{\mathrm{E}} \end{aligned}$$

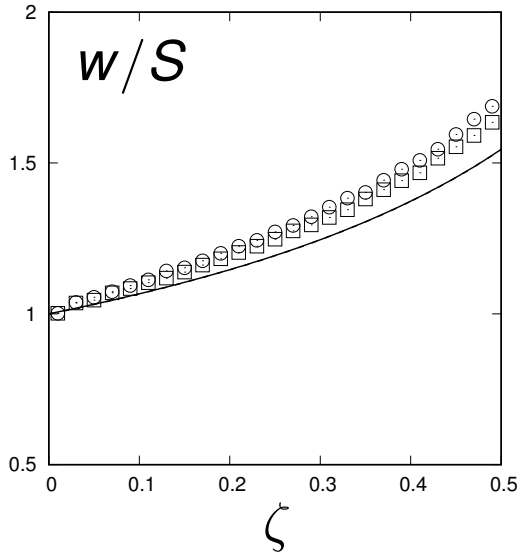
- numerical soln of coupled eqns for $(k, \rho, \tilde{u}, v, w)$ *completely indep* of true base hazard rate $\lambda_0(t)$



Variational theory: tests of order parameters



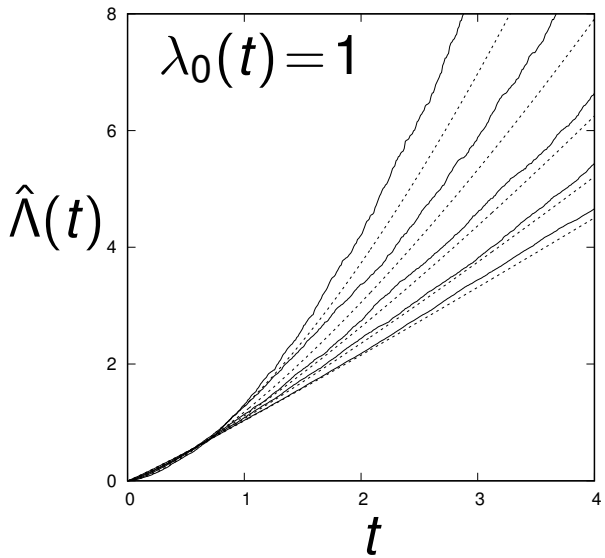


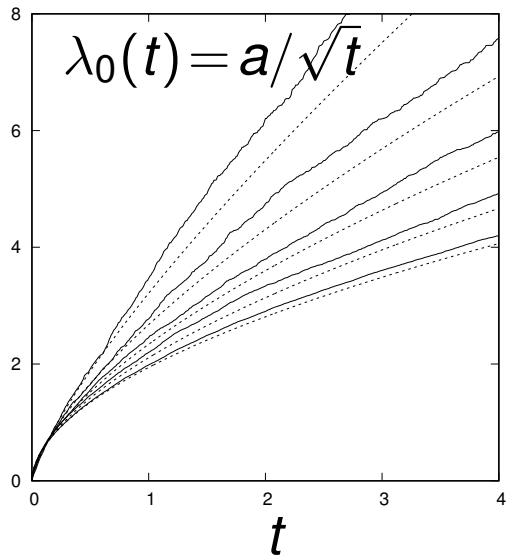


simulations: $N=200$, $\lambda_0(t)=1$, $S=0.5$ (circles) or $S=1$ (squares)

$N=400$, $S=0.5$,
a such that $\int \mathrm{d}t\, p(t)t = 1$,

$\zeta = 0.1, 0.2, 0.3, 0.4, 0.5$
(lower to upper curves)



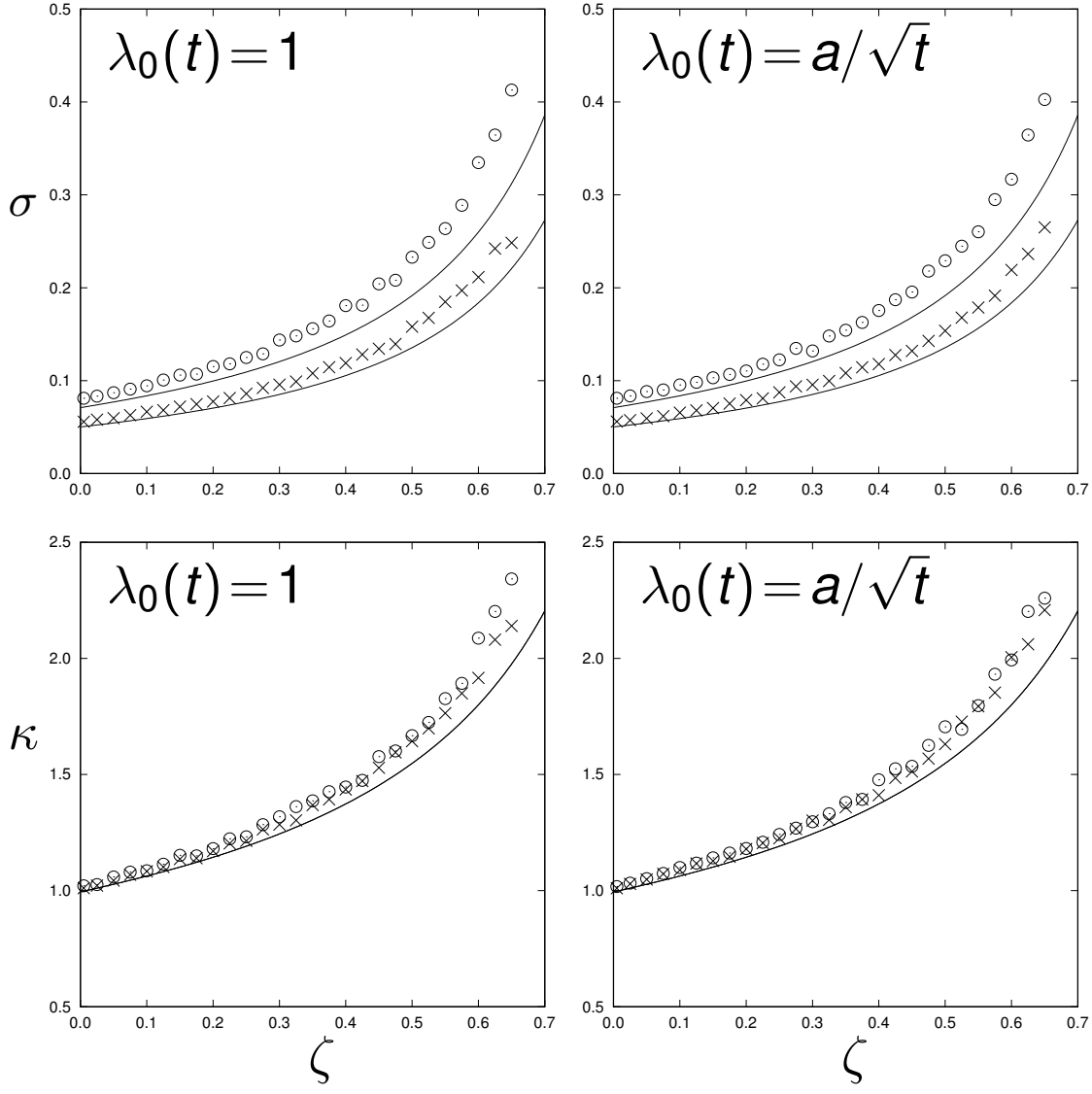


Variational theory: regression parameter clouds

slopes κ and widths σ of regression parameter clouds

$S = 0.5$,
$N = 200$ (circles)
or $N = 400$ (crosses)

solid lines:
variational theory



Summary

- Overfitting in Cox regression causes predictable bias

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

- Analytical approach based on statistical mechanics (replica method)

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

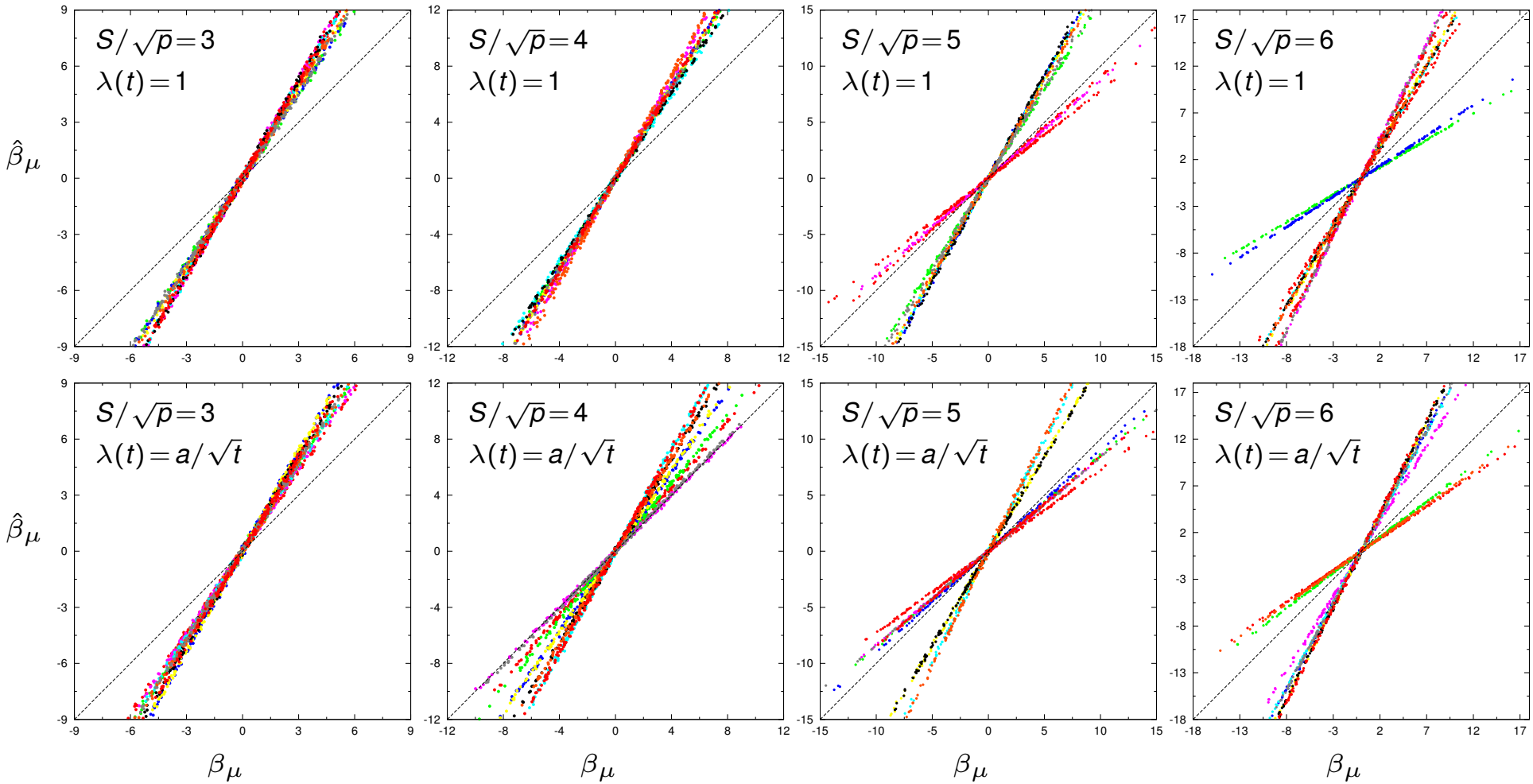
  predictions of variational theory: quite good,
  reliable basis for overfitting corrections

- Next
  - Include censoring ✓
  - Include priors/regularizers ✓

  - Analysis of exact equations (no variational approx)
  - Higher order variational approx
  - Associations for which $\sum_\mu \beta_\mu z_\mu$ is not Gaussian

  - Roll out overfitting correction protocols for Cox regression ✓
  - Other survival analysis models

Beyond replica symmetry



Large p/N and $\langle \beta_\mu^2 \rangle$: replica symmetry breaking

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

Censoring, competing risks and priors

- inclusion of censoring: essential for application to real data
- mathematically: censoring and multiple risks very similar
- inclusion of priors: MAP instead of ML regression

new set-up:

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

$r_i = 1$: *primary risk*
$r_i = 0$: *end of trial censoring*
$r_i > 1$: *other active risks*

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \mathrm{argmin}_{\boldsymbol{\theta}} \left[D(\hat{P}_{\mathcal{D}} || P_{\boldsymbol{\theta}}) - \frac{1}{N} \log p(\boldsymbol{\theta}) \right]$$

$$P(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} P(t_i, r_i | \boldsymbol{z}_i, \boldsymbol{\theta})$$

Multiple risk version of Cox model:

$$p(t, r | \boldsymbol{z}, \boldsymbol{\beta}, \lambda) = \lambda_r(t) \mathrm{e}^{\boldsymbol{\beta}^r \cdot \boldsymbol{z} - \sum_{r'=0}^{R} \exp(\boldsymbol{\beta}^{r'} \cdot \boldsymbol{z}) \Lambda_{r'}(t)}, \quad \Lambda_r(t) = \int_0^t \mathrm{d}s \, \lambda_r(s)$$

- prior: $p(\boldsymbol{\beta}) = Z^{-1} \prod_\mu \mathrm{e}^{-\phi(\beta_\mu/\kappa)}$

  theory will now involve covariate correlations,
  $A_{\mu\nu} = \langle z_\mu z_\nu \rangle - \langle z_\mu \rangle \langle z_\nu \rangle$

- more complex stat mech problem,
  but still doable:

$$E(\{\boldsymbol{\beta}^0, \lambda^0\}) = \lim_{\gamma\to\infty} \frac{\partial}{\partial\gamma} \lim_{n\to 0} \frac{1}{n} \mathrm{extr}_{\boldsymbol{\Xi},\boldsymbol{D},\mathbf{d},\lambda_{11},\ldots,\lambda_{Rn}} \Psi[\boldsymbol{\Xi}, \boldsymbol{D}, \mathbf{d}, \lambda_{11}, \ldots, \lambda_{Rn}]$$

$$\Psi[\ldots] = -\frac{1}{2}\zeta \sum_{rs=1}^{R} \Big[\sum_{\alpha\rho=1}^{n} D_{r\alpha,s\rho}\Xi_{r\alpha,s\rho} + \sum_{\alpha=1}^{n} \Big(d^a_{rs,\alpha}\Xi_{r\alpha,s0} + d^b_{rs,\alpha}\Xi_{r0,s\alpha} \Big) \Big]$$

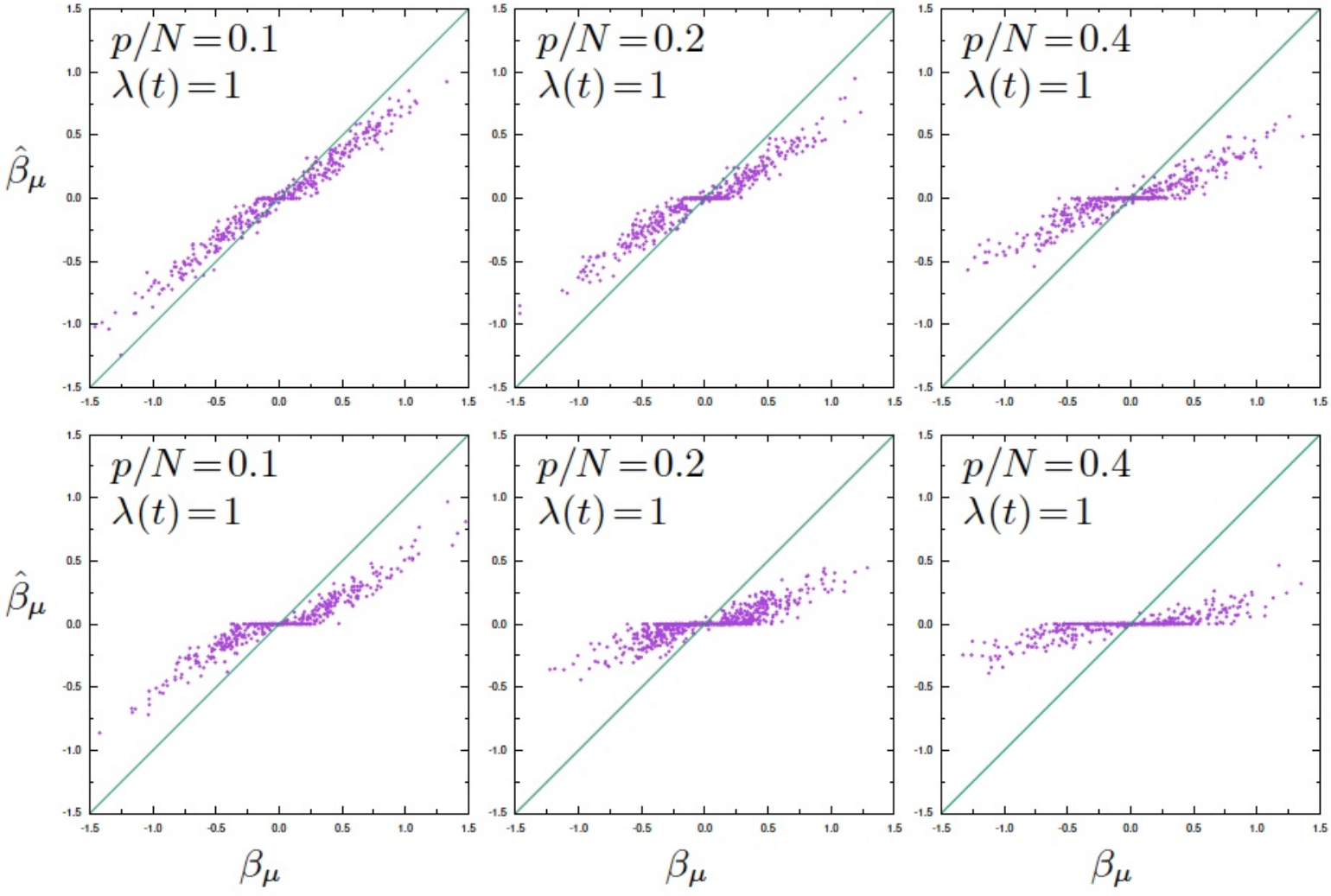
$$- \lim_{p\to\infty} \frac{\zeta}{p} \log \int \Big[\prod_{r=1}^{R} \prod_{\alpha=1}^{n} \mathrm{d}\boldsymbol{\eta}^{r\alpha} \mathrm{e}^{\gamma \sum_{\mu\leq p} [\phi(\eta^\star_\mu/\kappa) - \phi(\eta^{r\alpha}_\mu/\kappa)]} \Big]$$

$$\times \mathrm{e}^{\sum_{rs=1}^{R} \Big[-\mathrm{i} \sum_{\alpha\rho=1}^{n} \hat{\Xi}_{r\alpha,s\rho} \boldsymbol{\eta}^{r\alpha}.\boldsymbol{A}\boldsymbol{\eta}^{s\rho} - \mathrm{i} \sum_{\alpha=1}^{n} \left(\hat{\Xi}_{r\alpha,s0} \boldsymbol{\eta}^{r\alpha}.\boldsymbol{A}\boldsymbol{\eta}^{s0} + \hat{\Xi}_{r0,s\alpha} \boldsymbol{\eta}^{r0}.\boldsymbol{A}\boldsymbol{\eta}^{s\alpha} \right) \Big]}$$

$$- \log \int \mathrm{d}\boldsymbol{Y} \, \frac{\mathrm{e}^{-\frac{1}{2}\sum_{rs=1}^{R}\sum_{\alpha\rho=0}^{n} Y_{r\alpha}(\boldsymbol{\Xi}^{-1})_{r\alpha,s\rho} Y_{s\rho}}}{\sqrt{(2\pi)^{(n+1)R} \mathrm{Det}\boldsymbol{\Xi}}} \int \mathrm{d}t \sum_{r=0}^{R} p(t, r|\{y_0, \lambda_0\}) \prod_{\alpha=1}^{n} \Big[\frac{p(t, r|\{y_\alpha, \lambda_\alpha\})}{p(t, r|\{y_0, \lambda_0\})} \Big]^\gamma$$

non-Gaussian priors:
regression parameter clouds no longer linear

e.g. $p(\beta) = Z^{-1} \prod_\mu \mathrm{e}^{-|\beta_\mu|/\sigma}$



simplest case

Gaussian prior, $p(\boldsymbol{\beta}) = (\kappa/2\pi)^{p/2} \mathrm{e}^{-\frac{1}{2}\kappa\boldsymbol{\beta}^2}$,
only end-of-trial censoring

- only spectrum $\varrho(a)$ of covariate correlation matrix $\boldsymbol{A}$ appears in final RS equations
- explicit expression for $\lambda_0(t)$ of censoring risk
- two new scalar order parameters
- RS saddle point problem: extremize

$$\Psi_{\rm RS}[\ldots] = -\frac{1}{2}\zeta\Big[g(u^2+v^2+w^2) + \gamma\kappa S_0^2 - \langle \log(\gamma\kappa + ag)\rangle$$
$$- f\langle \frac{a}{\gamma\kappa+ag}\rangle - (\frac{wS_1}{S_0})^2 \langle \frac{a^2}{\gamma\kappa+ag}\rangle^{-1}\Big]$$
$$+ \gamma \int \mathrm{D}y_0 \int \mathrm{d}t \sum_{r=0}^{1} p(t,r|S_1 y_0, \lambda^\star) \log p(t,r|S_1 y_0, \lambda^\star)$$
$$- \int \mathrm{D}y_0 \mathrm{D}z \int \mathrm{d}t \sum_{r=0}^{1} p(t,r|S_1 y_0, \lambda^\star) \log \int \mathrm{D}y\; p^\gamma(t,r|uy+vz++wy_0,\lambda)$$

with: $S_\ell^2 = \frac{1}{p}\boldsymbol{\beta}^\star \boldsymbol{A}^\ell \boldsymbol{\beta}^\star$, $\langle g(a)\rangle = \int \mathrm{d}a\; \varrho(a) g(a)$

Overfitting in inference with max entropy 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:

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

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