Theory of overfitting in Cox regression

ACC Coolen, King’s College London London, Sept 13th 2017

Introduction
   Regression for time-to-event data
   Overfitting in Cox regression

Replica analysis of overfitting in PH regression
   The basic ideas
   Translation to Cox’s model
   Replica symmetric solution
   Resulting asymptotic prediction

Variational approximation

Tests and applications

Discussion
Introduction
   Regression for time-to-event data
   Overfitting in Cox regression

Replica analysis of overfitting in PH regression
   The basic ideas
   Translation to Cox’s model
   Replica symmetric solution
   Resulting asymptotic prediction

Variational approximation

Tests and applications

Discussion
Regression for time-to-event data

**Data**

\[ \mathcal{D} = \{(z_1, t_1), \ldots, (z_N, t_N)\} \]

samples \((z_i, t_i)\),
drawn indep from \(p(t, z)\)

\[ z_i \in \mathbb{R}^d : \text{d covariates (measured at } t = 0) \]

\[ t_i \in \mathbb{R}^+ : \text{failure time (death, onset of disease, ...)} \]

**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)

hazard rate: \( h(t|z) = \lambda(t) e^{\beta \cdot z} \)

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

dependent parameters: \( \beta = (\beta_1, \ldots, \beta_d) \), \( \lambda(t) \quad t \geq 0 \)

▶ Maximum Likelihood estimation

\[ (\hat{\beta}, \hat{\lambda}) = \underset{\beta, \lambda}{\operatorname{argmax}} \left\{ \frac{1}{N} \sum_i \log p(t_i|z_i, \beta, \lambda) \right\} \]

▶ Maximise over \( \lambda(t) \) first

\[ \hat{\lambda}(t|\beta) = \frac{\sum_j \delta(t-t_j)}{\sum_k \theta(t_k-t) e^{\beta \cdot z_k}} \quad \text{(Breslow estimator)} \]

\[ \hat{\beta} = \underset{\beta}{\operatorname{argmax}} \left\{ \sum_i \beta \cdot z_i - \sum_i \log \left[ \frac{\sum_j e^{\beta \cdot z_i} \theta(t_j-t_i)}{\sum_j \theta(t_j-t_i)} \right] \right\} \]
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 \( p(\mathcal{D}|\beta, \lambda) \))
  ▸ consistent base hazard rate, such that \( \int_0^\infty dt \lambda(t) = \infty \) (ML subject to constraint \( \int_0^\infty dt \lambda(t) = R \), then \( R \to \infty \))

▸ **Multiple risks**
  risk labels \( r_i \in \{0, \ldots, R\} \),
  \[ \mathcal{D} = \{(z_1, t_1, r_1), \ldots, (z_N, t_N, r_N)\} \]

▸ **Frailty, random effects and latent class models**
  (simple formulae only for special choices)
  \[
p(t|z, \beta, \lambda) = -\frac{d}{dt} \sum_{\ell=1}^L w_\ell \exp\left[-e^{\beta^\ell \cdot z} \int_0^t dt' \lambda^\ell(t')\right]
\]
What has changed since the 1970s?

▶ Medical data have evolved

▶ shear *volume* ...

▶ *diversity* of data sources
  (clinical, genomic, biomarkers, health records, imaging, …)

▶ *complexity* of experimental pipelines
  (confounders, batch effects, variability between centres, …)

▶ *dimension* mismatch
  then: ～500 samples, ～10 covariates
  now: ～1000 samples, ～10^6 covariates
Introduction
  Regression for time-to-event data
  Overfitting in Cox regression

Replica analysis of overfitting in PH regression
  The basic ideas
  Translation to Cox’s model
  Replica symmetric solution
  Resulting asymptotic prediction

Variational approximation

Tests and applications

Discussion
overfitting in Cox regression

ML method ...
p-values, z-scores, confidence intervals don’t measure overfitting!

rule of thumb: \( p_{\text{max}} = \text{events}/10 \)

- too optimistic ...
- must depend on \( \beta \) ...
- 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)

$\frac{p}{N} = 0.20$
\( N = 500, \)
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 $\beta$

- 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,

\[ \langle \beta^2 \rangle = 0.25 \]
Base hazard rates underestimated for short times, and over-estimated for large times ...

\[ \lambda(t) = 1 \]

\[ \lambda(t) = a/\sqrt{t} \]

\[ \hat{\lambda}(t) \]

\[ \hat{\lambda}(t) \]

\[ t \]

\[ t \]

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

Gaussian association pars, \( \langle \beta^2 \rangle = 0.25 \), \( N = 400 \), average event time \( \langle t \rangle = 1 \)
Intuition for the problem ...

- **Overfitting in ML regression**

  assumed model: $p_\theta$

  $\theta_{ML} = \arg\max_\theta p(\mathcal{D}|\theta) = \arg\min_\theta D(\hat{p}||p_\theta)$

  $\hat{p}(t, z) = \frac{1}{N} \sum_i \delta(t-t_i) \delta(z-z_i), \quad D(\hat{p}||p_\theta) = \int dt dz \hat{p}(t, z) \log \left[ \frac{\hat{p}(t|z)}{p(t|z, \theta)} \right]$

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

  true pars: $\theta^*$

  - fixed $d$: $\lim_{N \to \infty} \hat{p}(t, z) = p(t, z|\theta^*), \text{ so } \theta_{ML} = \theta^*$
  - $d = \mathcal{O}(N)$: $\lim_{N \to \infty} \hat{p}(t, z) \neq p(t, z|\theta^*)$ ...

- **Barrier to overfitting theory**

  want: study relation between $\theta_{ML}(\mathcal{D})$ and $\theta^*$, for $d = \mathcal{O}(N)$

  need: formula for $\theta_{ML}(\mathcal{D})$ ...
Introduction
Regression for time-to-event data
Overfitting in Cox regression

Replica analysis of overfitting in PH regression
The basic ideas
Translation to Cox’s model
Replica symmetric solution
Resulting asymptotic prediction

Variational approximation

Tests and applications

Discussion
The basic ideas

Step 1 – define a suitable overfitting measure

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

note that

\[
\theta_{\text{ML}} = \arg\min_{\theta} D(\hat{p}_{\theta^*} \| p_{\theta})
\]

\[
\theta = \theta^* : \quad D(\hat{p}_{\theta^*} \| p_{\theta}) = D(\hat{p}_{\theta^*} \| p_{\theta^*}) \quad \leftarrow \text{not zero!}
\]

Define:

\[
E(\theta^*, \mathcal{D}) = \min_{\theta} D(\hat{p}_{\theta^*} \| p_{\theta}) - D(\hat{p}_{\theta^*} \| p_{\theta^*})
\]

\[
E(\theta^*, \mathcal{D}) > 0 : \text{ underfitting}
\]

\[
E(\theta^*, \mathcal{D}) = 0 : \text{ optimal fitting}
\]

\[
E(\theta^*, \mathcal{D}) < 0 : \text{ overfitting}
\]

Typical behaviour

\[
E(\theta^*) = \left\langle E(\theta^*, \mathcal{D}) \right\rangle_{\mathcal{D}}
\]

\[
= \left\langle \min_{\theta} \left\{ \frac{1}{N} \sum_{i} \log \frac{p(t_i | z_i, \theta^*)}{p(t_i | z_i, \theta)} \right\} \right\rangle_{\mathcal{D}}
\]
Step 2 – eliminate minimisation over $\beta$

- **Laplace identity**  
  *(steepest descent)*

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

use in reverse:

$$
E(\theta^*) = \langle \min_{\theta} \left\{ \frac{1}{N} \sum_i \log \left[ \frac{p(t_i | z_i, \theta^*)}{p(t_i | z_i, \theta)} \right] \right\} \rangle_\mathcal{D}
$$

$$
= - \frac{1}{N} \langle \max_{\theta} \left\{ \sum_i \log \left[ \frac{p(t_i | z_i, \theta)}{p(t_i | z_i, \theta^*)} \right] \right\} \rangle_\mathcal{D}
$$

$$
= - \lim_{\gamma \to \infty} \frac{1}{N} \langle \frac{\partial}{\partial \gamma} \log \int \! d\theta \, e^{\gamma \sum_i \log \left[ \frac{p(t_i | z_i, \theta)}{p(t_i | z_i, \theta^*)} \right]} \rangle_\mathcal{D}
$$

$$
= - \lim_{\gamma \to \infty} \frac{1}{N} \frac{\partial}{\partial \gamma} \langle \log \int \! d\theta \, \prod_{i=1}^{N} \left[ \frac{p(t_i | z_i, \theta)}{p(t_i | z_i, \theta^*)} \right]^\gamma \rangle_\mathcal{D}
$$

interpretation:  
stochastic minimisation, with noise $\sim 1/\gamma$
Step 3 – enable averaging over $\mathcal{D}$

- **Replica method**
  \[
  \langle \log Z \rangle = \lim_{n \to 0} \frac{1}{n} \log \langle Z^n \rangle = \lim_{n \to 0} \frac{1}{n} \log \left\langle \prod_{\alpha=1}^{n} Z \right\rangle
  \]

  - evaluate for integer $n$,
  - analytical continuation to non-integer $n$

- **Application**
  \[
  E(\theta^*) = - \lim_{\gamma \to \infty} \frac{1}{N} \frac{\partial}{\partial \gamma} \left\langle \log \int d\theta \prod_{i=1}^{N} \left[ \frac{p(t_i|z_i, \theta)}{p(t_i|z_i, \theta^*)} \right]^\gamma \right\rangle_{\mathcal{D}}
  \]

  \[
  = - \lim_{\gamma \to \infty} \frac{1}{N} \frac{\partial}{\partial \gamma} \lim_{n \to 0} \frac{1}{n} \log \left\langle \left[ \int d\theta \prod_{i=1}^{N} \left[ \frac{p(t_i|z_i, \theta)}{p(t_i|z_i, \theta^*)} \right]^\gamma \right]^n \right\rangle_{\mathcal{D}}
  \]

  \[
  = - \lim_{\gamma \to \infty} \lim_{n \to 0} \frac{1}{Nn} \frac{\partial}{\partial \gamma} \log \int d\theta_1 \ldots d\theta^n \left\langle \prod_{i=1}^{N} \prod_{\alpha=1}^{n} \left[ \frac{p(t_i|z_i, \theta^\alpha)}{p(t_i|z_i, \theta^*)} \right] \right\rangle_{\mathcal{D}}
  \]

  \[
  = - \lim_{\gamma \to \infty} \lim_{n \to 0} \frac{1}{Nn} \frac{\partial}{\partial \gamma} \log \int d\theta_1 \ldots d\theta^n \left[ \int dz dt p(z)p(t|z, \theta^*) \prod_{\alpha=1}^{n} \left[ \frac{p(t|z, \theta^\alpha)}{p(t|z, \theta^*)} \right] \right]^N
  \]
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)
- *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, ...

\[
\frac{N}{d} \alpha_c(\kappa)
\]

Gardner theory for binary classifiers

massive overfitting
Introduction
Regression for time-to-event data
Overfitting in Cox regression

Replica analysis of overfitting in PH regression
The basic ideas
Translation to Cox’s model
Replica symmetric solution
Resulting asymptotic prediction

Variational approximation

Tests and applications

Discussion
Translation to Cox’s model

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

\[ \Lambda(t) = \int_0^t dt' \lambda(t') \]

▸ Defns, short-hands

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

\[ S^2 = \frac{1}{p}(\beta^*)^2, \quad \lambda^* = \lambda_0, \quad \alpha = \frac{d}{N} \]

▸ Insert, work out,

\begin{align*}
E(S, \lambda_0) &= -\lim_{\gamma \to \infty} \lim_{n \to 0} \frac{1}{n} \frac{\partial}{\partial \gamma} \extr_{\mathbf{c}, \lambda_1, \ldots, \lambda_n} \left\{ \frac{1}{2} \alpha n[1 + \log(2\pi)] + \frac{1}{2} \alpha \log \Det(C') \right. \\
& \quad + \log \int \frac{\mathbf{d}y \ e^{-\frac{1}{2} \mathbf{y} \cdot C^{-1} \mathbf{y}}}{\sqrt{(2\pi)^{n+1} \Det C}} \int dt \ p(t|y_0, \lambda_0) \prod_{\alpha=1}^n \left( \frac{p(t|y_0, \lambda_0)}{p(t|y_0, \lambda_0)} \right)^\gamma \left. \right\}
\end{align*}

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

\[ C': \quad n \times n, \quad C'_{ab} = C_{ab} - C_{a0}C_{0b} / C_{00}^2, \quad a, b = 1 \ldots n \]
Introduction
  Regression for time-to-event data
  Overfitting in Cox regression

Replica analysis of overfitting in PH regression
  The basic ideas
  Translation to Cox’s model
  Replica symmetric solution
  Resulting asymptotic prediction

Variational approximation

Tests and applications

Discussion
Replica symmetric solution

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

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

interpretation:

\[
c_0 = \lim_{p \to \infty} \frac{1}{p} \beta^* \cdot \langle \langle \beta \rangle \rangle_D, \quad c = \lim_{p \to \infty} \frac{1}{p} \langle \langle \beta^2 \rangle \rangle_D, \quad C = \lim_{p \to \infty} \frac{1}{p} \langle \langle \beta^2 \rangle \rangle_D
\]

Insert into formulae,
diagonalise \( C \) and \( C' \), manipulations, integrations,
take the limit \( n \to 0 \) ...
take the limit \( \gamma \to \infty \) ...
Introduction
  Regression for time-to-event data
  Overfitting in Cox regression

Replica analysis of overfitting in PH regression
  The basic ideas
  Translation to Cox’s model
  Replica symmetric solution
  Resulting asymptotic prediction

Variational approximation

Tests and applications

Discussion
Resulting asymptotic prediction

\[ E(S, \lambda_0) = \int dt \, p(t) \log \left[ \frac{\lambda_0(t)}{\lambda(t)} \right] - \]

\[ (1 + \ddot{u}^2) \left[ 1 - \frac{1}{\ddot{u}^2} \int Dz Dy_0 \int dt \, p(t | S_0, \lambda_0) W \left( \ddot{u}^2 e^{\ddot{u}^2 + wy_0 + vz \Lambda(t)} \right) \right] \]

\[ Dz = (2\pi)^{-1/2} e^{-\frac{1}{2} z^2} \, dz, \]

\[ W(z): \text{Lambert } W\text{-function}, \]

\[ \ddot{u}, v, w, \lambda(t) \text{ to be solved from} \]

\[ \zeta v^2 = \int Dz Dy_0 \int dt \, p(t | S_0, \lambda_0) \left[ \ddot{u}^2 - W(\ddot{u}^2 e^{\ddot{u}^2 + wy_0 + vz \Lambda(t)}) \right]^2 \]

\[ \zeta = \int Dz Dy_0 \int dt \, p(t | S_0, \lambda_0) \frac{W(\ddot{u}^2 e^{\ddot{u}^2 + wy_0 + vz \Lambda(t)})}{1 + W(\ddot{u}^2 e^{\ddot{u}^2 + wy_0 + vz \Lambda(t)})} \]

\[ 0 = \int Dz Dy_0 \, y_0 \int dt \, p(t | S_0, \lambda_0) W(\ddot{u}^2 e^{\ddot{u}^2 + wy_0 + vz \Lambda(t)}) \]

\[ \frac{p(t)}{\lambda(t)} = \int Dz Dy_0 \int_t^\infty dt' \, p(t' | S_0, \lambda_0) \frac{W(\ddot{u}^2 e^{\ddot{u}^2 + wy_0 + vz \Lambda(t')})}{\ddot{u}^2 \Lambda(t')} \]

\[ \square \]
interpretation:

\[ v^2 = \frac{1}{p} \left\{ \langle \beta^2 \rangle_D - \left( \frac{\beta^*}{|\beta^*|} \cdot \langle \beta \rangle_D \right)^2 \right\}, \quad w = \frac{1}{\sqrt{p}} \frac{\langle \beta \rangle_D \cdot \beta^*}{|\beta^*|} \]

link with data clouds:

- slope: \( \kappa = w/S \)
- width: \( \sigma = v/\sqrt{p} \)

special limits for \( \zeta = p/N \):

- \( \zeta \to 0 \): no overfitting, \( v \to 0, \ w \to S, \ \lambda(t) \to \lambda_0(t) \)
- \( \zeta \to 1 \): phase transition, \( v, w \to \infty \)

main numerical challenge:

\[
\frac{p(t)}{\lambda(t)} = \int DzDy_0 \int_t^\infty dt' \ p(t' | Sy_0, \lambda_0) \frac{W(\tilde{u}^2 e^{\tilde{v}^2 + wy_0 + vz} \Lambda(t'))}{\tilde{u}^2 \Lambda(t')} \]

\[ t \gg 1 : \ \log \Lambda(t) = \rho \log \Lambda_0(t) + (1 - \rho) \log(\log \Lambda_0(t)) + \ldots \]

\[ \rho = \frac{w}{2S} \left( 1 + \sqrt{1 + 4\tilde{u}^2/w^2} \right) \]
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 $\sigma$ and slope $\kappa$ of data clouds for $S = 0.5$ and $\langle t \rangle = 1$

lines: variational theory

$\circ: N = 200$

$x: N = 400$
overfitting correction of inflated association parameters using slope predicted by variational theory.
$N = 200$, $p = 80$, $\langle \beta^2 \rangle = 0.25$
\[ \hat{\Lambda}(t) = 1 \quad \hat{\Lambda}(t) = a/\sqrt{t} \]

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 $\langle \beta^2 \rangle$: replica symmetry breaking (disconnected solution spaces)

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

- Overfitting in Cox regression causes predictable bias
  - (i) inflation of association parameters
  - (ii) hazard rates: underestimated \((t \text{ small})\), overestimated \((t \text{ large})\)

- 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!