Self-consistency as a method to develop computationally effective algorithms for high-dimensional models

Alex Tsodikov

Department of Biostatistics
School of Public Health

University of Michigan

A collection of thoughts and examples

- Complex nonlinear models with high-dimensional parameters
  - Mechanistic models formulated as average over unobserved structural complete-data model

  - Statistical models that are difficult to fit because of dimensionality

- Algorithms
  - Algorithms based on self-consistency
    - EM algorithms working through missing data imputation
    - Generalizations that are not based on missing data (MM, etc.)

High-dimensional models

- Functional parameters whose dimension is proportional to sample size (non-parametrically specified distributional characteristics)
- Big data sets (cancer registry data, large population trials)
- Survival analysis, categorical data analysis, multivariate response models

M and Z-estimation algorithms

- Target function (loglikelihood ℓ)
- Model parameters ω
- Estimating equation $\varphi(\omega, \omega) = 0$
- Estimation algorithm
  - Nonlinear programming

    $$\varphi(\omega^{(k+1)}, \omega^{(k)}) = 0$$

    $$\omega^{(k+1)} = \psi(\omega^{(k)})$$

  - Convergence
    - Fixed point $\omega^{(k+1)} \to \omega^{(k)} \to \hat{\omega}$
    - Contraction mapping $\|\psi\| < 1$

Transforms

- Are used to simplify solutions to difficult problems

Original problem

Transform

Simpler problem

Solution

Inverse transform

Laplace transform

$$\mathcal{L}_X(s) = \mathbb{E}\{e^{-s\cdot X}\} = \int_0^\infty e^{-sx} f(x)\,dx$$

Reversed approach

- Recognize original problem as a transform of a simpler one

Transform

$\omega, \tilde{U}$

Original problem

Simpler problem

?

Solution

ω

Self-consistency

The EM framework: solve (1) by solving (2)

- MLE problem

$$\ell(\omega) = \log L(\omega)$$

model parameters

(1) $\max_{\omega} \ell(\omega)$

- Transform

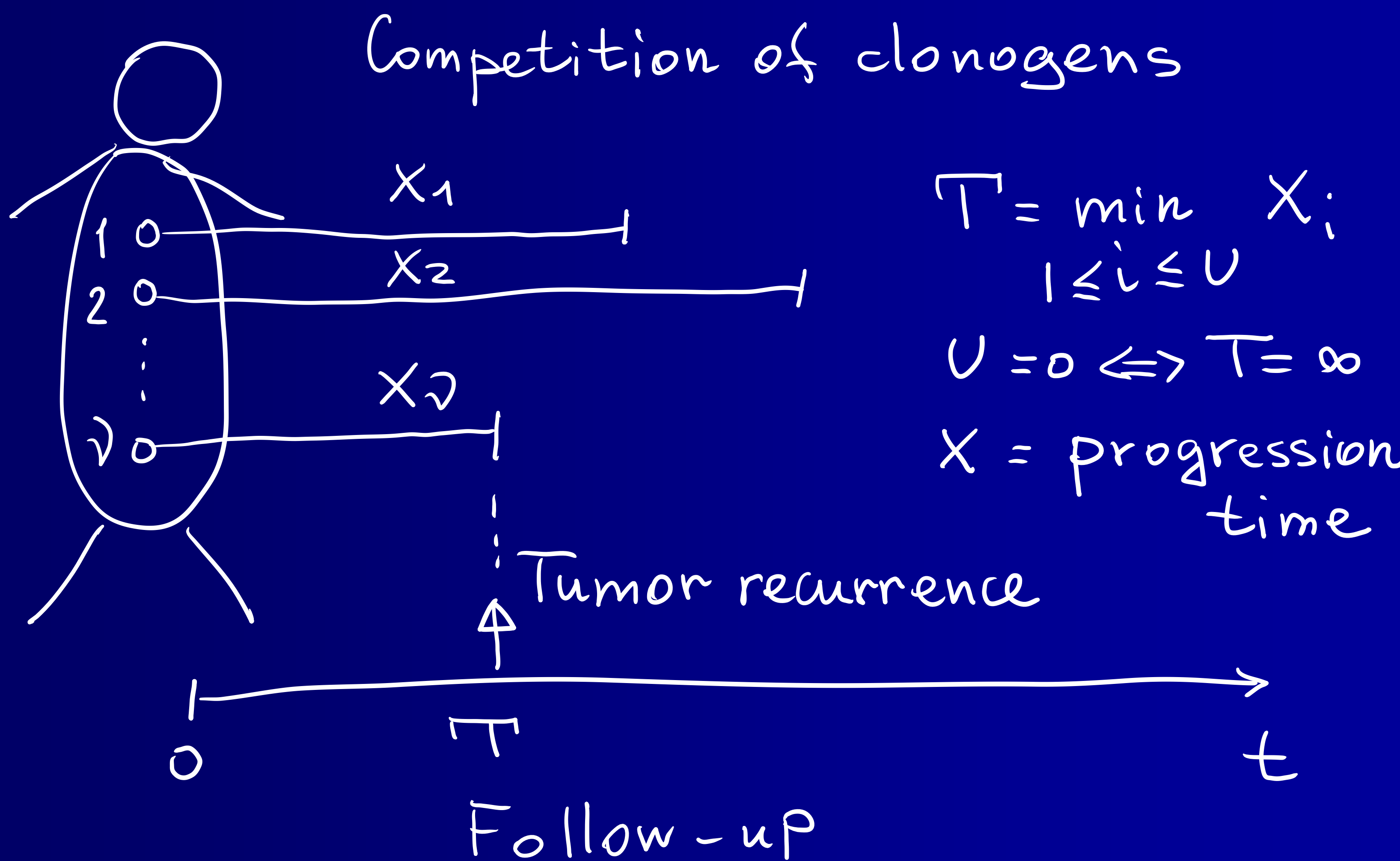
$$L(\omega) = \mathbb{E}\{L_0(\omega, \upsilon)\}$$

missing data

(2) $\max_{\omega} \ell_0(\omega, \upsilon)$

$$\ell_0 = \log L_0$$

A Simple Model of Tumor Recurrence



Distribution of T when U is allowed to vary?

Survival function $G(t) = \Pr\{T > t\}$

$F(t) = \Pr\{X > t\}$

$X = \text{i.i.d.}$

$$G(t \mid U) = [F(t)]^U$$

$$G(t) = \mathbb{E}\{G(t \mid U)\} = e^{-\Theta[1 - F(t)]}$$

$U \sim \text{Poisson}(\Theta)$

count

A cure model

- Covariates affect progression time

$$G(t) = e^{-\theta(z)\left[1 - \underbrace{F(t)^{\eta(z)}}_{\text{Nested Cox model for X}}\right]}$$

- Formulation without missing data
  - For the Cure cumulative hazard must be bounded => model as const*CDF

Univariate Frailty Models

$$G(t \mid z) = \mathbb{E}\{F^{U}(t) \mid z\} = \gamma(F) = \mathcal{L}(H)$$

$$U \sim \mathbb{P}(du \mid \theta(z), \eta(z), \overset{?}{\ldots})$$

γ is probability generating function

$\mathcal{L}$ is Laplace transform of U

H is baseline cumulative hazard

$$H = -\log F$$

Self-consistency algorithm

- Covariates affect progression time

$$G(t) = e^{-\theta(z)\left[1 - \underbrace{F_{(t)}^{\eta(z)}}_{\text{Nested Cox model for } X}\right]}$$

- Formulation without missing data
  - For the Cure cumulative hazard must be bounded => model as const*CDF

Univariate survival MLE

- Loglikelihood

$$\ell = \sum_t D_t \cdot \log \Delta H_t + \sum_t \text{Smth}\left(H[0,t]\right)$$

Baseline cumulative hazard

censoring

- Score equation

$$\frac{\partial \ell}{\partial \Delta H_\tau} = \frac{D_\tau}{\Delta H_\tau} - \sum_{\text{at Risk}} \delta \, \Theta(H)$$

Self-consistency algorithm, univariate survival

jump of cumulative hazard @ t

number of failures @ t

$$\Delta H_{\tau}^{(m+1)} = \frac{D_{\tau}}{\sum_{@Risk} \Theta\left(H^{(m)} \mid z, \delta\right)}$$

δ - censoring index $= \begin{cases} 1, & \text{failure} \\ 0, & \text{cens} \end{cases}$

Univariate frailty model

- Laplace transform

$U \geq 0$, r.v. $\qquad \mathcal{L}(s \mid z)$

$\mathbb{P}(du \mid z)$

s.f. $G(t \mid z) = \mathcal{L}(H \mid z)$

$H(t)$ = cumulative hazard

- The model

$$G(t \mid z) = \mathbb{E}\{\underbrace{e^{-U \cdot H(t)}}_{\text{PH model}}\}$$

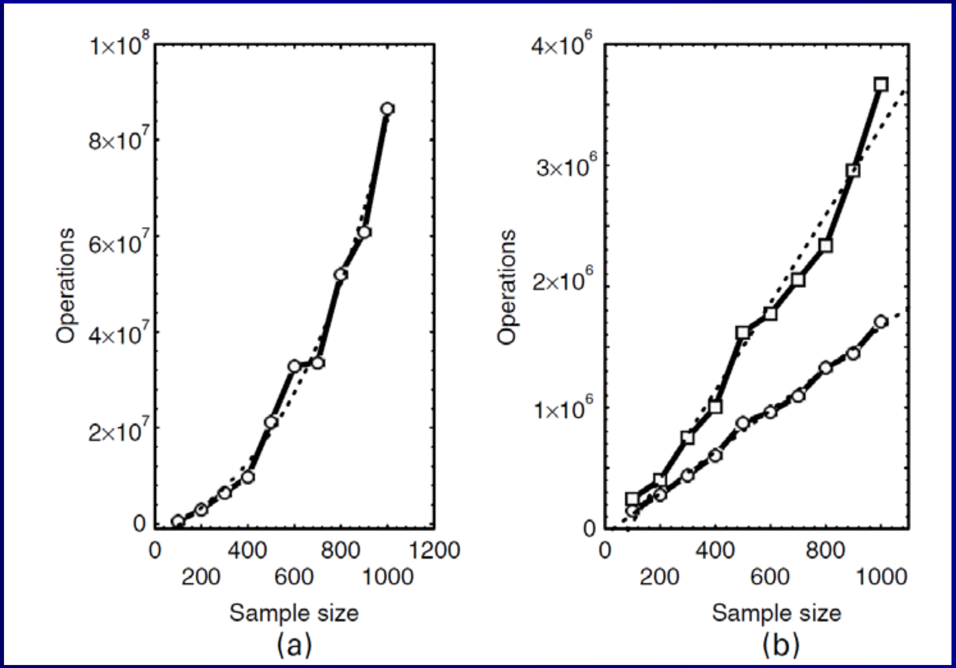
Imputation operator

$$\Theta(H \mid z, \delta) = -\frac{\mathcal{L}^{(\delta+1)}(H|z)}{\mathcal{L}^{(\delta)}(H|z)}$$

$$\delta = \begin{cases} 1, & \text{failure} \\ 0, & \text{censoring} \end{cases}$$

$$= \mathbb{E}\{ U \mid \underbrace{(t, z, \delta)}_{\text{Observed data}} \}$$

Performance of the algorithm



Full MLE
Nonlinear programming

Self-consistency
algorithm

nltm package for R implements the algorithm for a variety of survival models, tried on registry data with hundreds of thousands patients

PHPH Model: Prostate Cancer Dose Escalation

Table 1. Estimates of the probability of cure and 95% likelihood ratio confidence intervals (in parenthesis) as estimated using stratified parametric analysis (SPA) based on (6) and multivariate semiparametric regression analysis (MSRA) based on (8).

		Dose group			
Prognostic category	Analysis	1	2	3	4
Favourable	SPA	0.80	0.74	0.87	1.00[a]
		(0.59, 0.93)	(0.61, 0.85)	(0.79, 0.94)	–
	MSRA	0.78	0.79	0.88	1.00[a]
		(0.55, 0.95)	(0.68, 0.92)	(0.80, 0.97)	–
Intermediate	SPA	0.25	0.51	0.58	0.74
		(0.12, 0.42)	(0.41, 0.61)	(0.48, 0.68)	(0.61, 0.85)
	MSRA	0.37	0.53	0.67	0.79
		(0.21, 0.55)	(0.41, 0.64)	(0.58, 0.78)	(0.68, 0.87)
Unfavourable	SPA	0.00	0.27	0.33	0.64
		(0.00, 0.02)	(0.16, 0.34)	(0.26, 0.42)	(0.53, 0.75)
	MSRA	0.02	0.35	0.46	0.60
		(0.00, 0.11)	(0.25, 0.45)	(0.38, 0.56)	(0.45, 0.74)

[a]The estimate of the probability of cure is set to be equal to 1 because there are no failures observed in this group of patients.

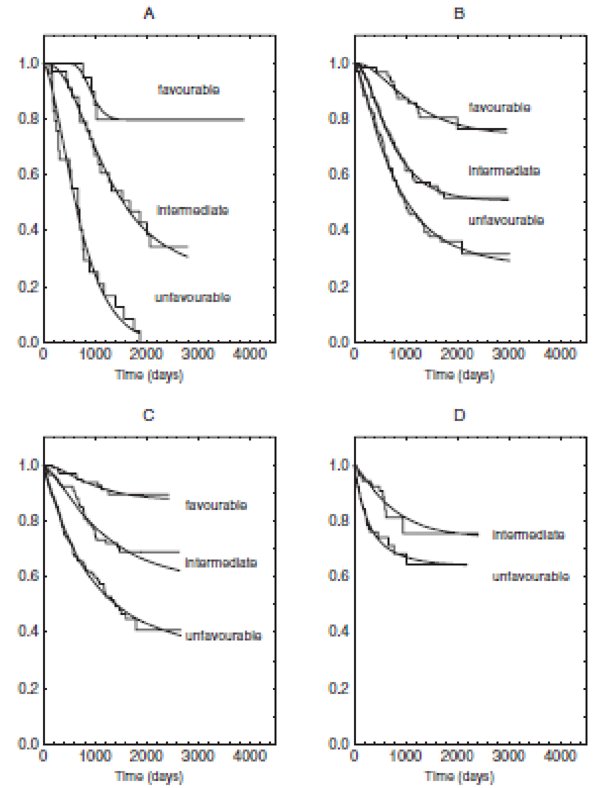


Figure 1. Survivor functions for relapse-free survival in different dose groups (A: group 1, B: group 2, C: group 3, D: group 4) of patients with clinically localized prostate cancer. Stepwise curve—Kaplan–Meier estimate, solid line—maximum likelihood parametric estimate based on formula (6).

Example: multinomial model

- Distribution of the response conditional on covariates

$$p_k = \Pr\{Z = k\} = \frac{\theta_k}{\sum_{j=1}^{N} \theta_j}$$

$$\theta_k = e^{\beta_k^T x}$$

Restriction $\theta_1 := 1, \quad \beta_1 := 0$

Artificial Mixture Transform

- Write the model as a quasi-mixture

$$p_k = \Theta_k \cdot \mathbb{E}\left\{ e^{-u \sum_{j=2}^{N} \Theta_j} \right\} = \frac{\Theta_k}{\sum_{j=1}^{N} \Theta_j}$$

- Is not really a mixture since is not a probability

Complete-data "likelihood"

- Poisson likelihood with an offset

fake missing data

"events" "person-yrs" "rate"

$$\sum_{\alpha} \sum_{i} I_{i\alpha} \cdot \log \Theta_{i\alpha} - u_i \cdot \Theta_{i\alpha}$$

$$I_{i\alpha} = \begin{cases} 1, & i\text{th subject response is } \alpha \\ 0, & \text{otherwise} \end{cases}$$

factorize

i = Subject

α = Category of response

Imputation: E-Step

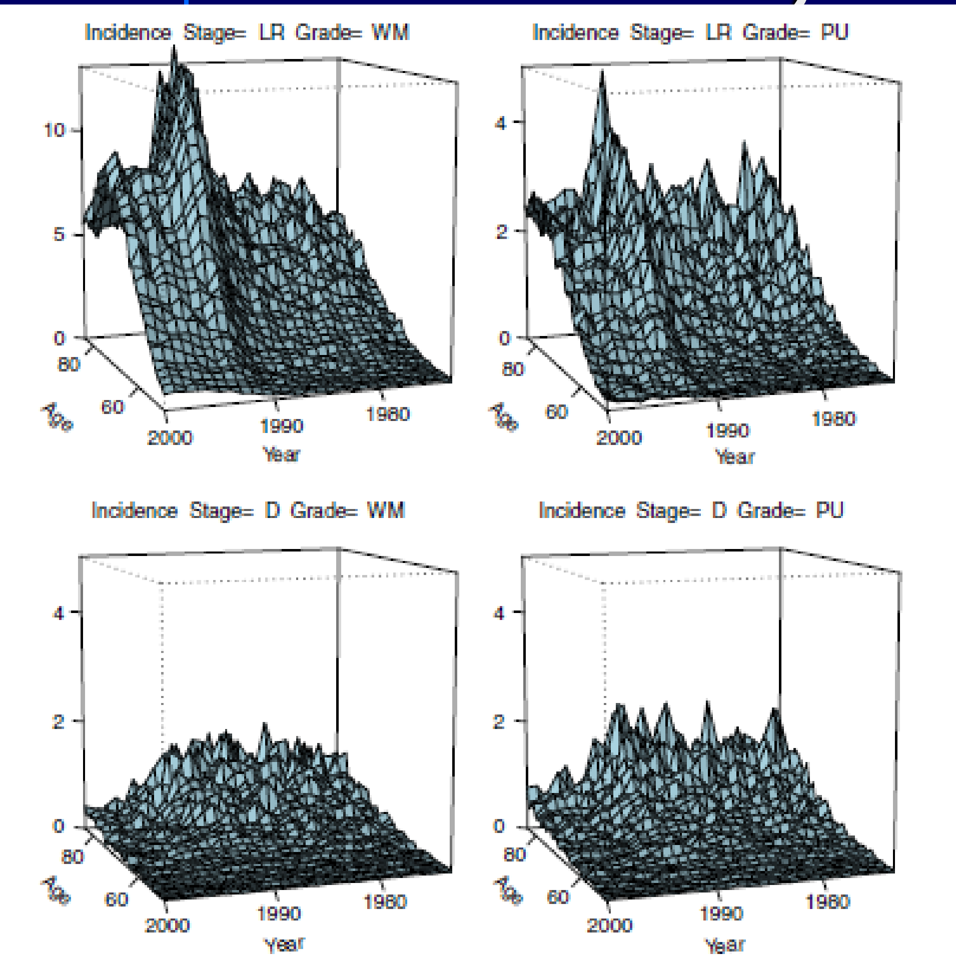
iteration

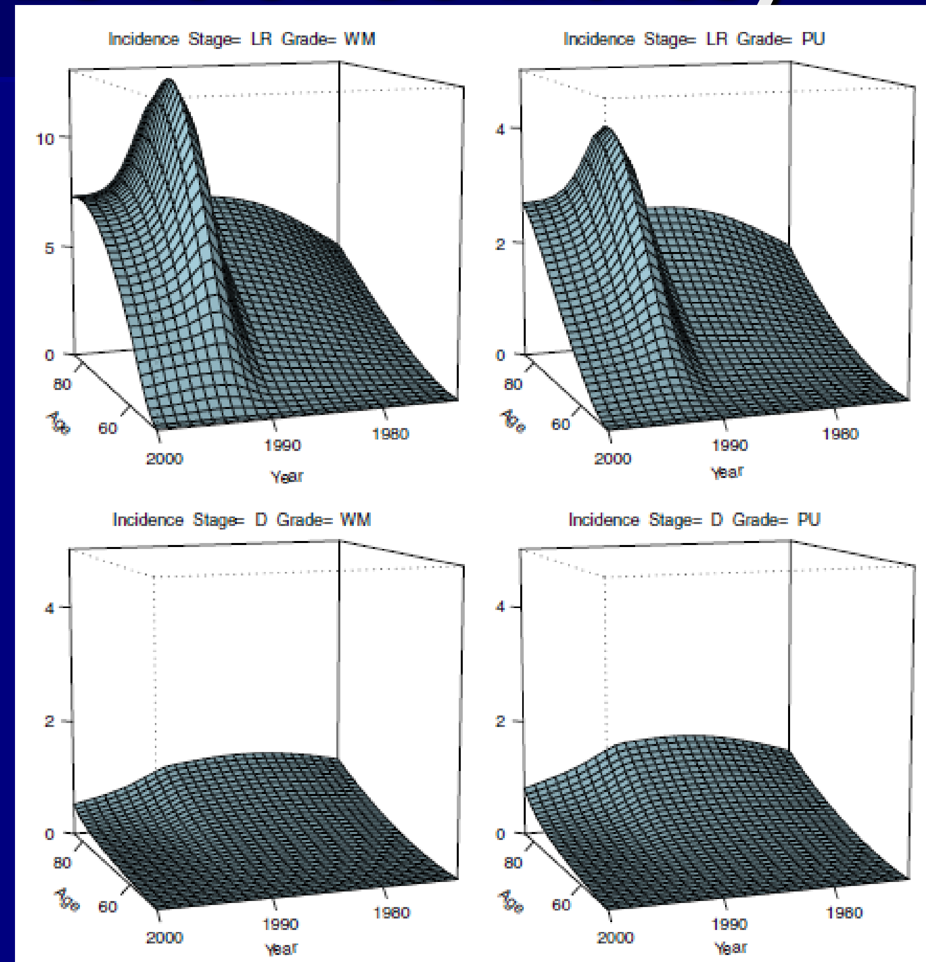
$$\hat{u}_i^{(m)} = \frac{1}{1 + \sum_{k=2}^{N} \Theta_{ik}} \quad \text{category}$$

$$\Theta_{ik} = e^{\beta_k^{(m)T} x_i}$$

Subject

Prostate Cancer Incidence SEER registry data (2-500,000 cancer cases, 11-15% of US males)





Chefo, S. and **Tsodikov, A**. (2009). Stage-specific cancer incidence: an artificially mixed multinomial logit model. *Statistics in Medicine*, 28/15, 2054-2076

Further examples and applications

- Multivariate (clustered) survival data

  Tsodikov, A., Liu, L., and Tseng, C. (2019) Likelihood Transformations and Artificial Mixtures, In *Statistical Modeling for Biological Systems,* Almudevar A, Oakes D, Hall J. Eds. Springer, in press.

- Missing data as a stochastic process, dynamic frailty

  Rice, J., **Tsodikov, A**. (2017) Semiparametric Time-to-Event Modeling in the Presence of a Latent Progression Event, ***Biometrics***, 73/2, 463-472.

Multivariate survival
Shared frailty model
Archimedian Copula models

- Transform

$U \geq 0$, r. v., Shared frailty

$\mathcal{L}(s \mid z)$

- The model

$$G(t_1, \ldots, t_n) = \mathcal{L}(H_1 + \ldots + H_n \mid z)$$

$$= \mathbb{E}\left\{ \underbrace{e^{-U \cdot \sum_{i=1}^{n} H_i}}_{\text{Independent PH models}} \right\}$$

Imputation operator

$$\Theta(H_* \mid z, n) = -\frac{\mathcal{L}^{(\delta_*+1)}(H_* \mid z)}{\mathcal{L}^{(\delta_*)}(H_* \mid z)}$$

$$\delta_* = \sum_{i=1}^{n} \delta_i \quad , \quad \delta_i = \begin{cases} 1, & \text{failure} \\ 0, & \text{censoring} \end{cases}$$

$$H_* = H_1 + \ldots + H_n$$

Multivariate survival
M-Step

- Symmetric case: Shared cumulative hazard

$$H_1 = H(t_1), \; \ldots, \; H_n = H(t_n)$$

$$\Delta H_m = \frac{D_m}{\sum_{s \in R_m} \Theta(H_*^s \mid z, \delta_*^s)}$$

δ_*^s: # failures in cluster s

R_m: Set of clusters at risk

$$H_*^s = \sum_i H(t_i^s)$$

t_i^s: times in cluster s

Archimedian Copula Models

- Shared frailty model induces an Archimedian Copula

marginal s.f.

$$G(t_1,\dots,t_n) = \mathcal{L}(H_*), \quad H_* = \sum H_i$$

$$\to M_i(t_i) = G(0,\dots,0,t_i,0,\dots,0) = \mathcal{L}(H_i)$$

$$H_i = \mathcal{L}^{-1}(M_i)$$

$$G = \mathcal{L}\left(\mathcal{L}^{-1}(M_1) + \dots + \mathcal{L}^{-1}(M_n)\right)$$

Shared frailty model vs. Archimedian Copula models

- Frailty models is a subset of Copula models
- Copula generator does not have to be a Laplace transform

n - Monotonic $(-1)^k \mathcal{L}^{(k)} > 0, \; k = 1, \ldots, n$

- $\mathcal{L}$ is a Laplace transform iff
  Completely monotonic

$(-1)^n \mathcal{L}^{(n)} \geq 0$

$n = 0, 1, \ldots, \infty$

- Archimedian Copula model serving clusters of any size is a shared frailty model

Characterization of positive dependence

- Multivariate totally positive Copulas of order 2, MTP2

$(-1)^n \mathcal{L}^{(n)}$ is log-convex

i.e. $\Theta(H \mid n) = -\dfrac{\mathcal{L}^{(n+1)}(H)}{\mathcal{L}^{(n)}(H)}$

is decreasing

Monotonic convergence of the algorithm

- Iterations

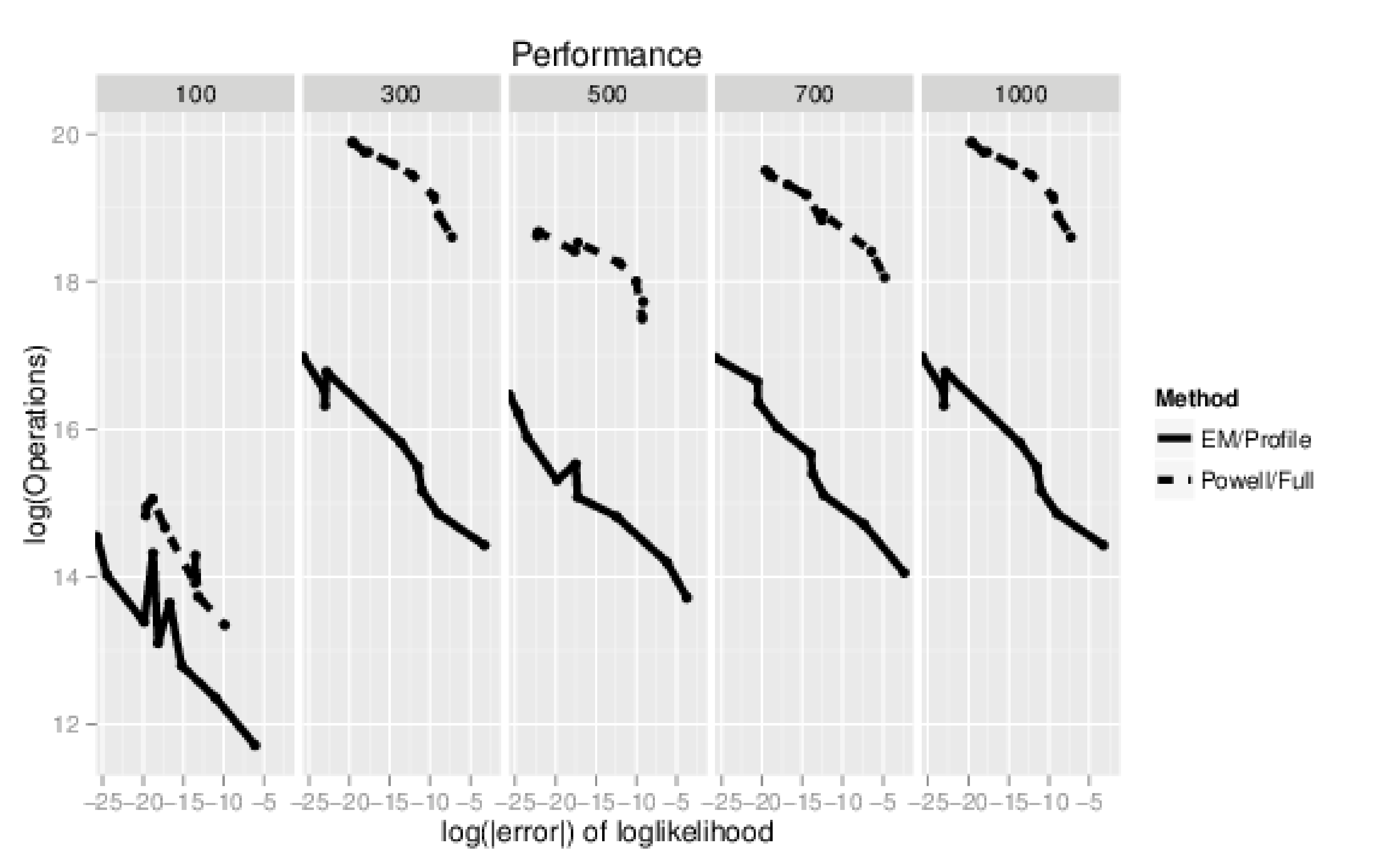
$$\Delta H_m = \frac{D_m}{\sum_{s \in R_m} \Theta(H_*^s \mid z, \delta_*^s)}$$

next iteration

previous iteration

- If Archimedian Copula is MTP2, each iteration improves the likelihood

Performance of the algorithm for clustered survival data



Latent progression event

- The model

Latent event model

$$d\Lambda_0(t|\mathbf{z}) = \lim_{h\to 0}\frac{P(T_0\in[t,t+h)|T_0\geq t,\mathbf{z})}{h} = \mu\, dH(t)$$

Model for observed event given latent

$$d\Lambda_1(t|T_0,\mathbf{z}) = \lim_{h\to 0}\frac{P(T_1\in[t,t+h)|T_1\geq t,T_0,\mathbf{z})}{h} = (t > T_0)\eta\, dH(t).$$

The self-consistency algorithm

$$0 = \sum_{i=1}^{n} \frac{dN_i(s)}{dH_0^{(k+1)}(s)} - \Psi_i^{(k)}(s) + \left[\frac{dH_0{}^{(k)}(s)}{dH_0^{(k+1)}(s)} - 1\right]\theta_i^{(k)}(s),$$

Observed failures Imputed latent failures

$$dH^{(k+1)}(s) = \frac{\sum_{i=1}^{n} dN_i(s) + \left[\sum_{i=1}^{n} \theta_i^{(k)}(s)\right] dH^{(k)}(s)}{\sum_{i=1}^{n} \left[\Psi_i^{(k)}(s) + \theta_i^{(k)}(s)\right]}.$$

Imputed latent risk set

$$\Psi_i^{(k)}(s) = Y_i(s) \frac{\eta^{1-\Delta_i} \mu e^{-\mu H^{(k)}(T_i^*)} - \eta \mu^{1-\Delta_i} e^{-\eta H^{(k)}(T_i^*)}}{\eta^{1-\Delta_i} e^{-\mu H^{(k)}(T_i^*)} - \mu^{1-\Delta_i} e^{-\eta H^{(k)}(T_i^*)}}$$

$$\theta_i^{(k)}(s) = (\eta - \mu)\mu^{1-\Delta_i} \frac{Y_i(s) e^{-\eta H^{(k)}(T_i^*) + (\eta-\mu) H^{(k)}(s)} + (1-\Delta_i)[1 - Y_i(s)] e^{-\mu H^{(k)}(s)}}{\eta^{1-\Delta_i} e^{-\mu H^{(k)}(T_i^*)} - \mu^{1-\Delta_i} e^{-\eta H^{(k)}(T_i^*)}}.$$