Bayesian Semi-parametric Biostatistics

Wes Johnson
UC Irvine

Dirichlet Process

- For $G \sim DP(c, F_0)$

$$G(\cdot) = \sum_{h=1}^{\infty} p_h \delta_{\theta_h}(\cdot)$$

$$\theta_h \overset{iid}{\sim} F_0$$

$$w_h \overset{iid}{\sim} Beta(1, c) \quad p_h = w_h \prod_{i=1}^{h-1} (1 - w_i)$$

  $\sum_h p_h = 1$ with probability one
- G is discrete with probability one

Dirichlet Process Mixture

We say X is drawn from a DPM if:

$$\begin{aligned} X \mid \theta &\sim G_\theta \\ \theta \mid P &\sim P \\ P \mid F_0, c &\sim DP(c, F_0) \end{aligned}$$

or

$$X|P \sim \int G_\theta(\cdot) P(d\theta) = \sum_{h=1}^{\infty} p_h G_{\theta_h}(\cdot)$$

$$\theta_h \overset{\text{iid}}{\sim} F_0$$

Polya Trees

- Split sample space Ω into two disjoint sets B_0 and B_1; further split B_0 into B_{00} etc:

B_0		B_1	
B_{00}	B_{01}	B_{10}	B_{11}

- $$Y_0 = P(X \in B_0), \quad Y_1 = P(X \in B_1),$$
$$Y_{00} = P(X \in B_{00} | X \in B_0),$$
$$Y_{01} = P(X \in B_{01} | X \in B_0),$$
$$Y_{10} = P(X \in B_{10} | X \in B_1),$$
$$Y_{11} = P(X \in B_{11} | X \in B_1).$$

- Then $P(X \in B_{ij}) = Y_i Y_{ij}$

- Let $\epsilon = \epsilon_1 \cdots \epsilon_m$ be an arbitrary binary number of dimension m

- Split $B_\epsilon \rightarrow \{B_{\epsilon 0}, B_{\epsilon 1}\} \qquad \forall \epsilon$.

- Then

$$\left.\begin{aligned} Y_{\epsilon 0} &= P(X \in B_{\epsilon 0} | X \in B_\epsilon) \\ Y_{\epsilon 1} &= P(X \in B_{\epsilon 1} | X \in B_\epsilon) \end{aligned}\right\} \Rightarrow$$

$$P(X \in B_{\epsilon_1 \cdots \epsilon_m}) = \prod_{j=1}^{m} Y_{\epsilon_1 \cdots \epsilon_j}$$

PT

- Random PM for G:

$$(Y_{\epsilon 0}, Y_{\epsilon 1}) \sim \text{Beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$$

- Center on F_0 by selecting the partition sets to be appropriate quantiles of F_0
- Let $\alpha_\epsilon = cm^2$ at level $m, \forall m$ (results in abs cont G w/ prob 1)
- We say $G|F_0, c \sim PT(c, F_0), \qquad E(G(\cdot)) = F_0(\cdot)$
- Finite Polya Tree is truncated at say level M
- Large c results in a parametric analysis, and small c results in a more non-parametric analysis

Mixture of Finite PTs

- Center on parametric family $\{F_\theta,\ \theta \in \Theta\}$
- Prior on θ, $p(\theta)$
- We say $G|F_\theta, c \sim PT(c, F_\theta), \qquad E(G(\cdot)) = F_\theta(\cdot)$, or

$$G \sim \int PT(c, F_\theta) p(d\theta)$$

- Truncated at level M results in an MFPT
- Large c results in analysis based on the parametric family

Bayesian Nonparametric and Semiparametric Inference for Disease Risk, ROC Curves, and Prevalence

Adam Branscum
Wes Johnson
Tim Hanson
Ian Gardner

Statistics in Medicine, 2008

Common Epidemiologic Problems

- Quantify the discriminatory ability of diagnostic screening test when applied to diseased and non-diseased individuals
- Obtain predictive probability of disease (PPD) for given test outcomes
- Estimate the prevalence of disease in a population

Setting

- Random sample of *n* individuals from a population with disease prevalence π
- Apply a *continuous* diagnostic test to each sampled individual
- 'Serology scores' measure concentration of serum antibodies specific to an antigen or agent
- Large values associated with diseased individuals
- Let Y denote the transformed serology score $Y = h(S)$

Setting

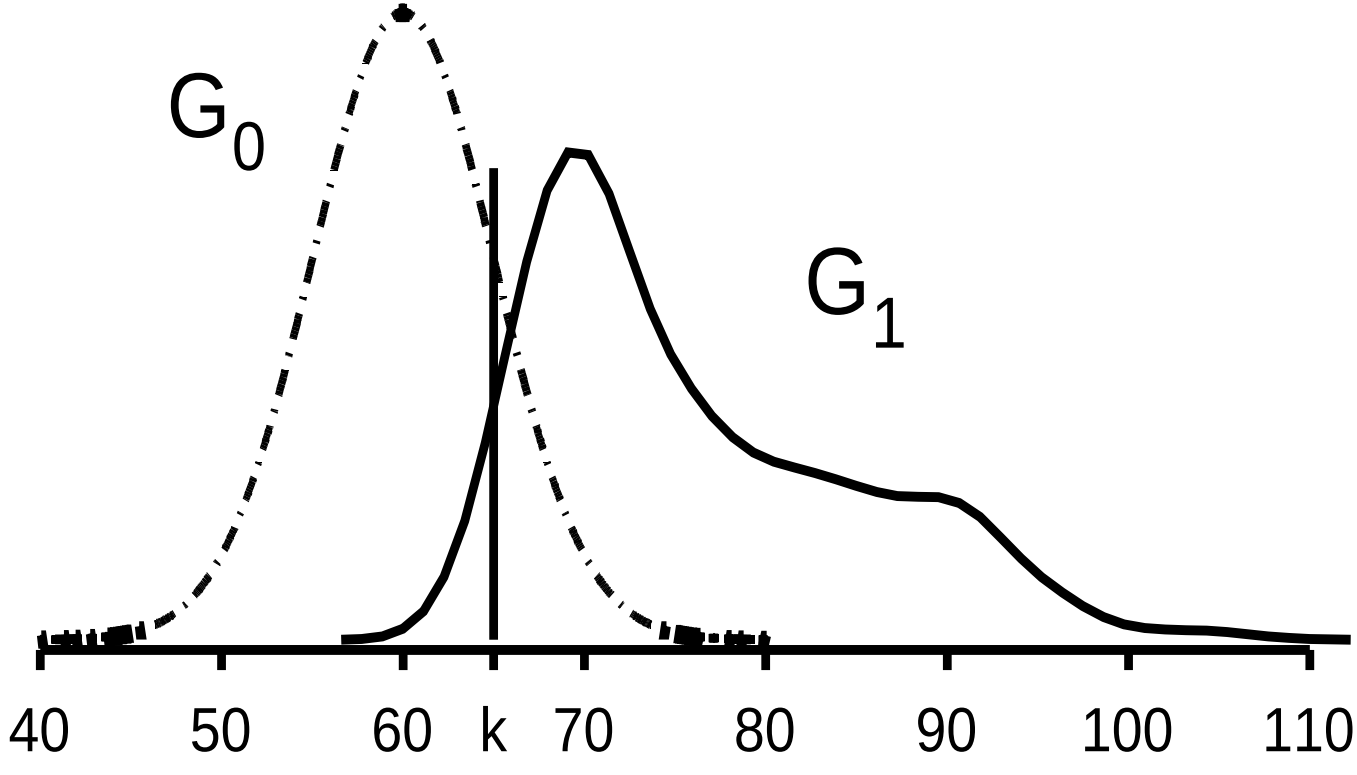
- Let D^+ and D^- denote disease positive and negative
- Distribution of serology scores

$$Y|D^- \sim G_0, \quad Y|D^+ \sim G_1$$

- True disease status is *unknown*
- Referred to as no gold-standard data

Standard Approach

- Dichotomize the serology scores using a cutoff value k



Standard Approach

- If $Y_i \geq k$ then classified as T^+, else, T^-

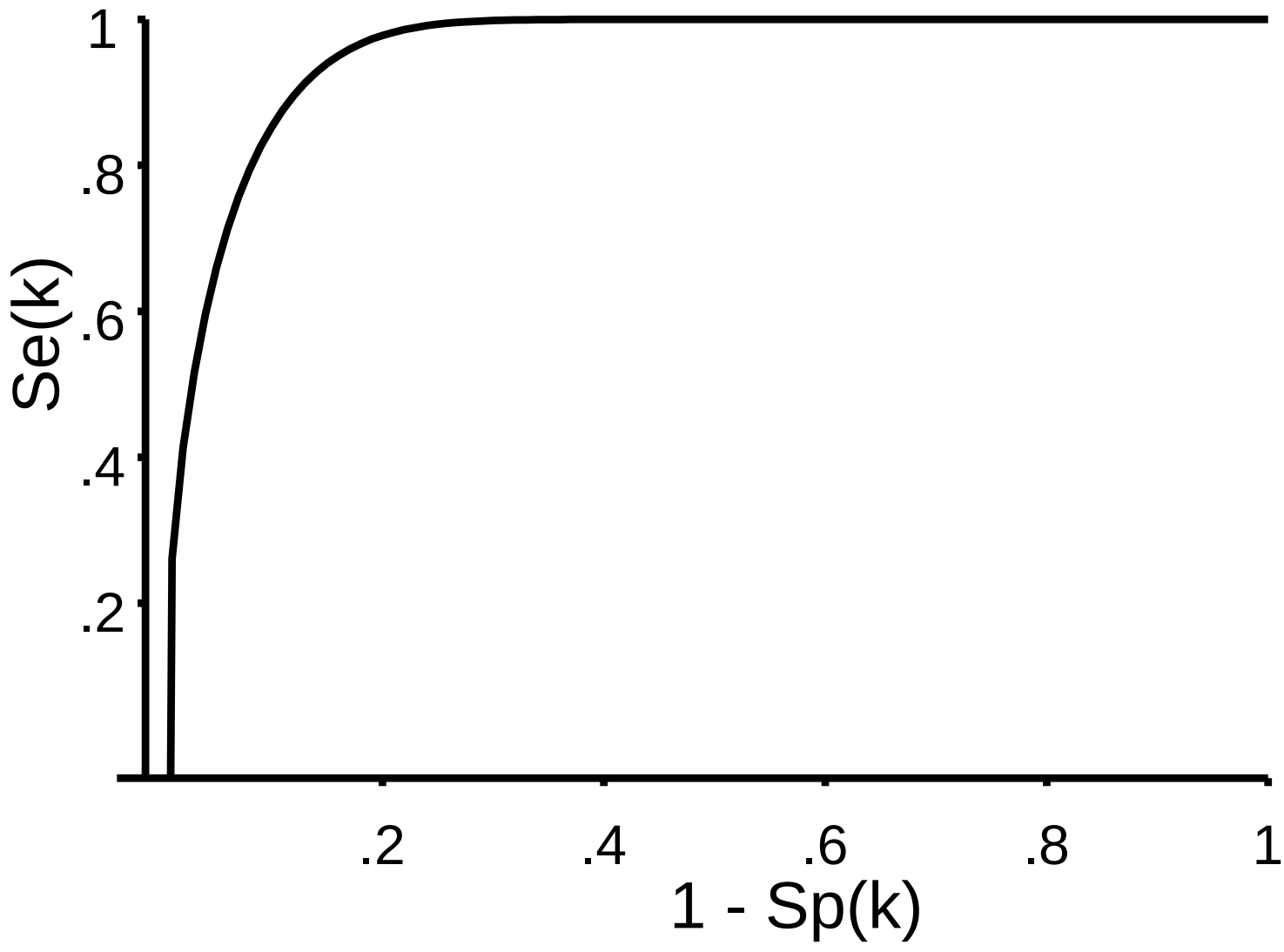
- $$Se = \mathsf{Pr}(T^+|D^+) = \mathsf{Pr}(Y > k|D^+) = 1 - G_1(k)$$

$$Sp = \mathsf{Pr}(T^-|D^-) = \mathsf{Pr}(Y < k|D^-) = G_0(k)$$

Receiver Operating Characteristic Curves

- Obtain sensitivity and specificity across *all* possible cutoff values
- The ROC curve plots $([1 - Sp(k)], Se(k))$ for all cutoff values *k* used to dichotomize the data
- Then select k that gives some form of “optimal" tradeoff between having high sensitivity and low false positive rate

ROC Curve



Modeling Serology Scores

- Let Z_i denote the latent binary indicator of disease status where $Z_i = 1$ if subject i is diseased

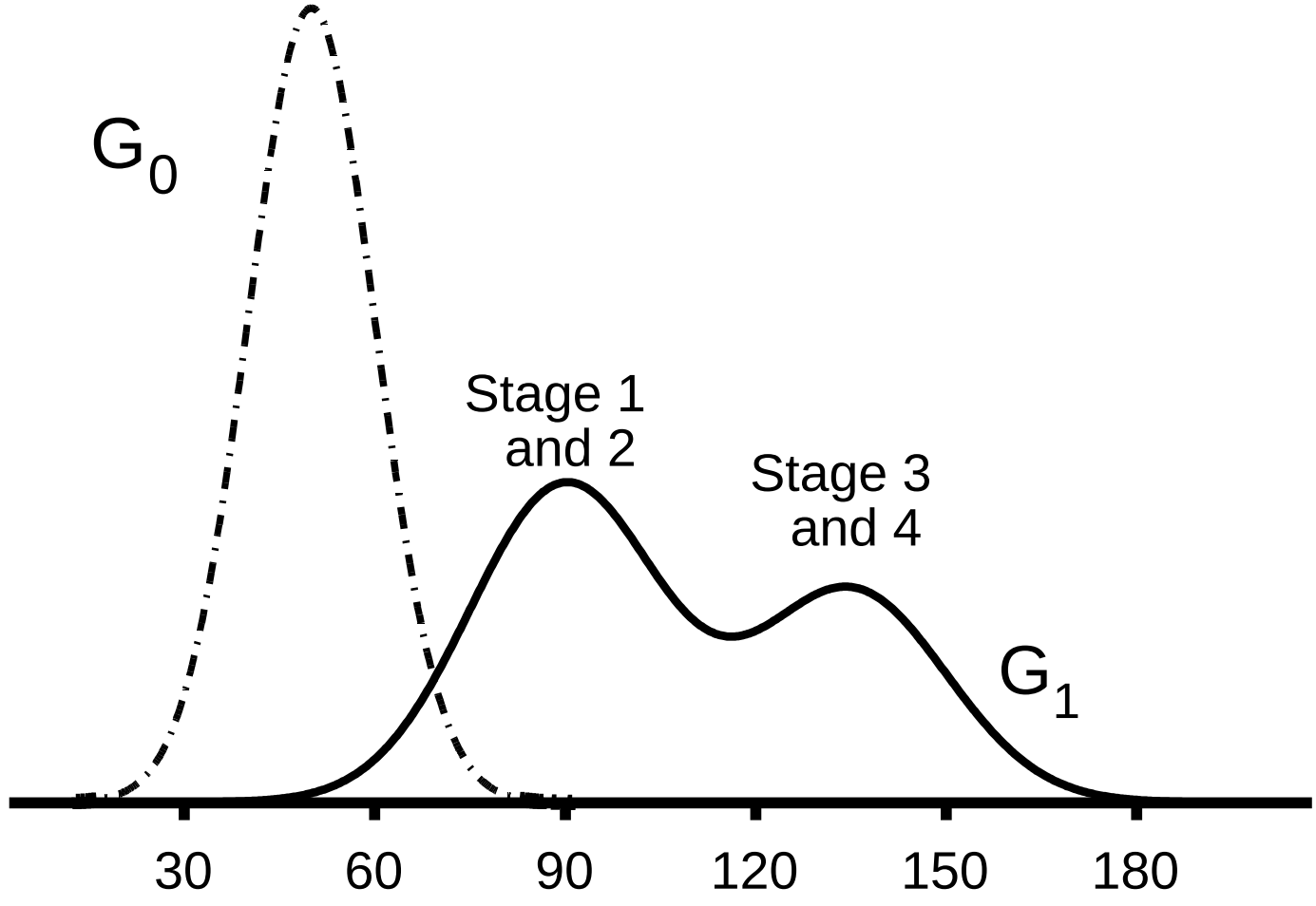
- The general model is

$$\begin{aligned} Y_i \mid G_0 &\sim G_0, \quad Z_i = 0 \\ Y_i \mid G_1 &\sim G_1, \quad Z_i = 1 \\ Z_i \mid \pi &\sim \text{Bern}(\pi) \end{aligned}$$

- How to specify flexible models for G_0 and G_1 when we don't know who is diseased and who is not?

Modeling Issues

- Serology scores for diseased individuals are generally coming from a mixture distribution based on staging



Nonparametric Model

- Use mixtures of Finite Polya Trees for Gs (Hanson and Johnson, 2002; Hanson, 2006)
- Centered on a parametric family like log normal, so we generalize the bi-normal
- Model lacks identifiability
- Solution
  - Incorporate subjective prior information
  - Incorporate covariates related to disease status
  - The latter may play role of “surrogate" gold standard

Semi-parametric Extension

$$
\begin{aligned}
Y_i \mid G_0 &\sim G_0, \quad Z_i = 0 \\
Y_i \mid G_1 &\sim G_1, \quad Z_i = 1
\end{aligned}
$$

Nonparametric		Semiparametric
π	$\rightarrow$	π_i
$Z_i \mid \pi \overset{iid}{\sim} \text{Bern}(\pi)$	$\rightarrow$	$Z_i \mid \pi_i \overset{\perp}{\sim} \text{Bern}(\pi_i)$
		$\pi_i = F(x_i\beta)$

Extension

- $F(w) = \frac{e^w}{1+e^w}$ yields logistic regression
- Get ROC curves
- Get covariate-specific predictive inferences for the probability of disease for a subject with serology score y and covariate x:

$$\Pr(Z = 1 | \, y, x, \text{data})$$

Illustration with NGS

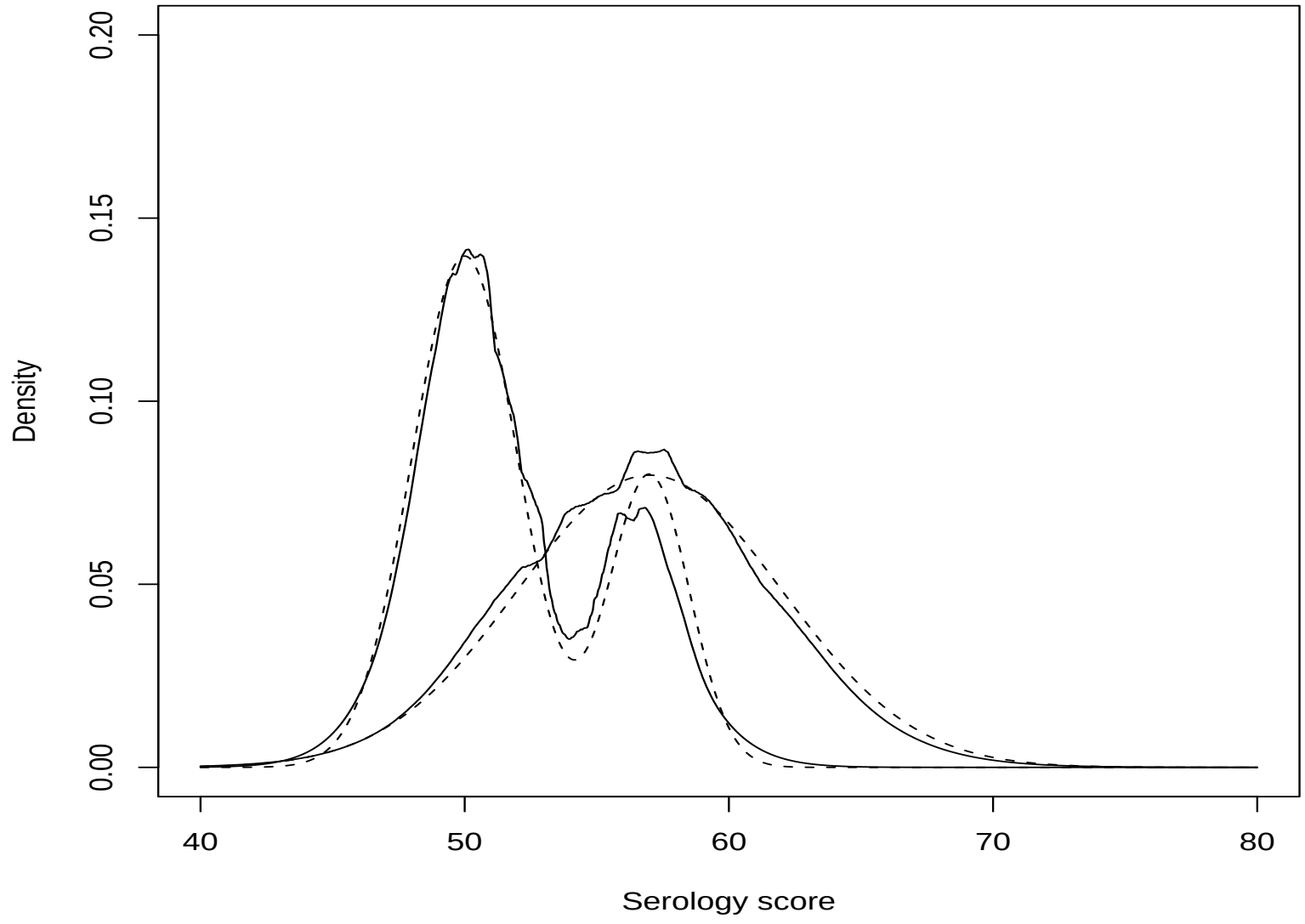
- Situation where distributions G_0 and G_1 have large overlap
- W/O additional information pertaining to true disease status, a *gold-standard* analysis is required.
- Suppose a binary covariate, x, that is associated with disease status is available.
- Believed that

$$\pi_1 = Pr(D^+|x = 1) = 0.95,\ \pi_0 = 0.05.$$

Use indep beta priors where

$$Pr(\pi_1 \geq 0.90) = 0.95 = Pr(\pi_0 \leq 0.1)$$

Illustration: Density estimates (solid lines)



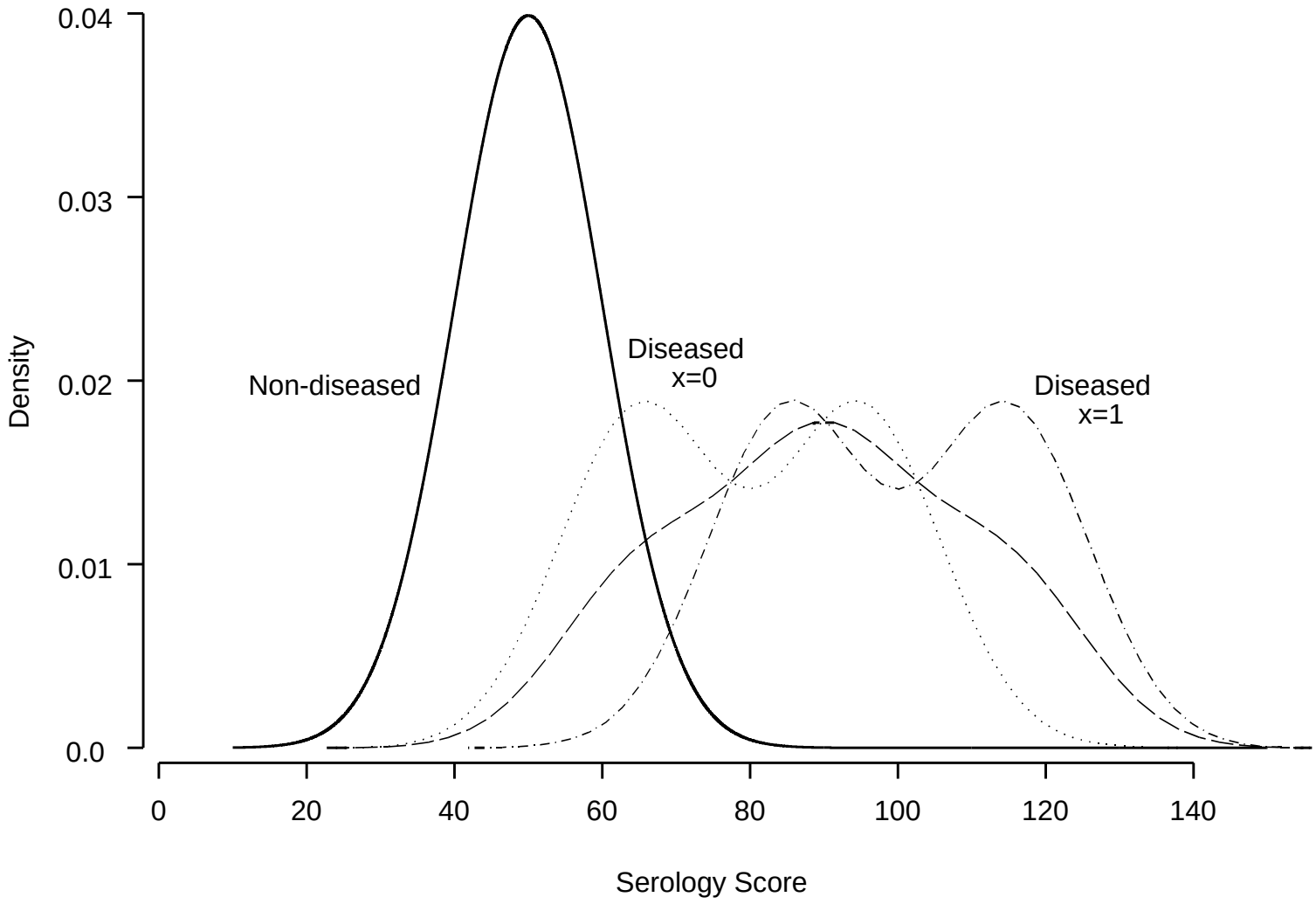
Further Extension

- Covariates related to disease status and serology score
- For diseased individuals,

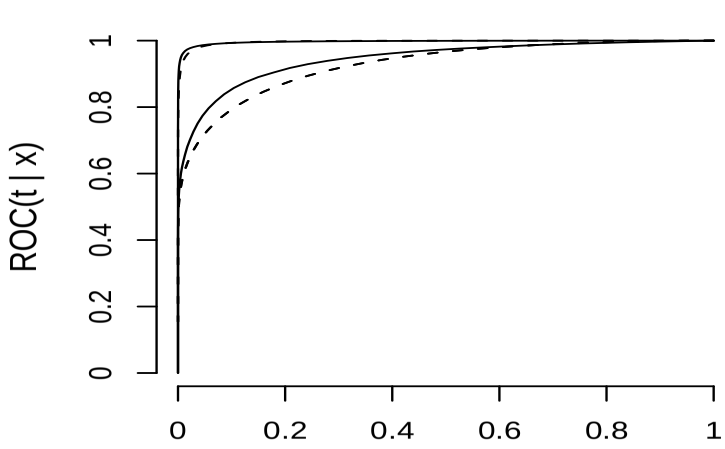
$$\begin{aligned} \log(Y_i) &= x_i^* \alpha + \epsilon_i \\ \epsilon_i | G_1 &\sim G_1 \\ G_1 | \sigma_1^2 &\sim FPT(N(0, \sigma_1^2), c_1) \end{aligned}$$

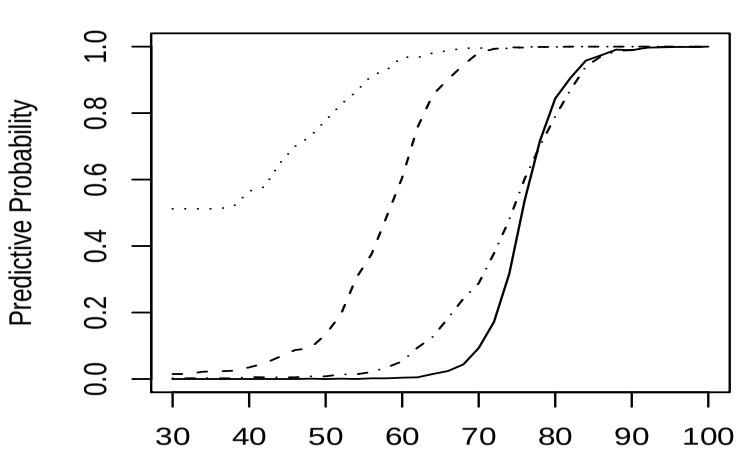
- BCJ prior on α
- Model for non-diseased individuals not expected to depend on covariates

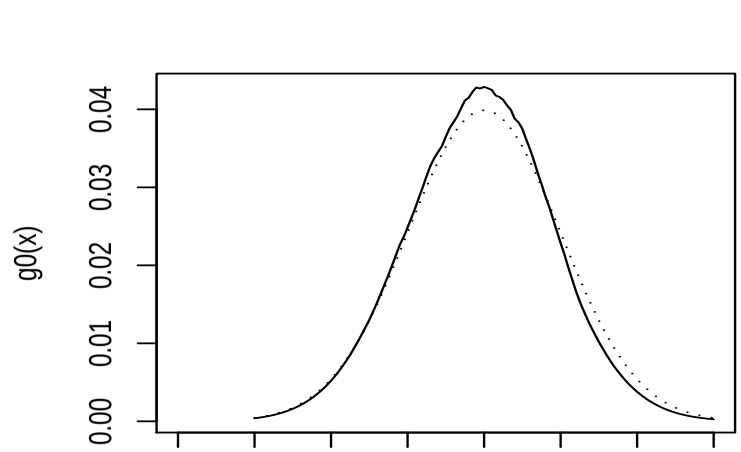
Distns of Y: (non-D, $x = 0$ $x = 1$)

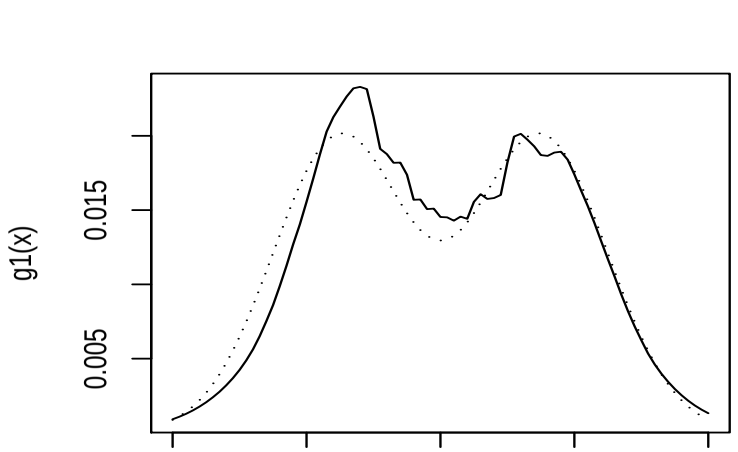


Est ROC and PP Dis : x=0,1







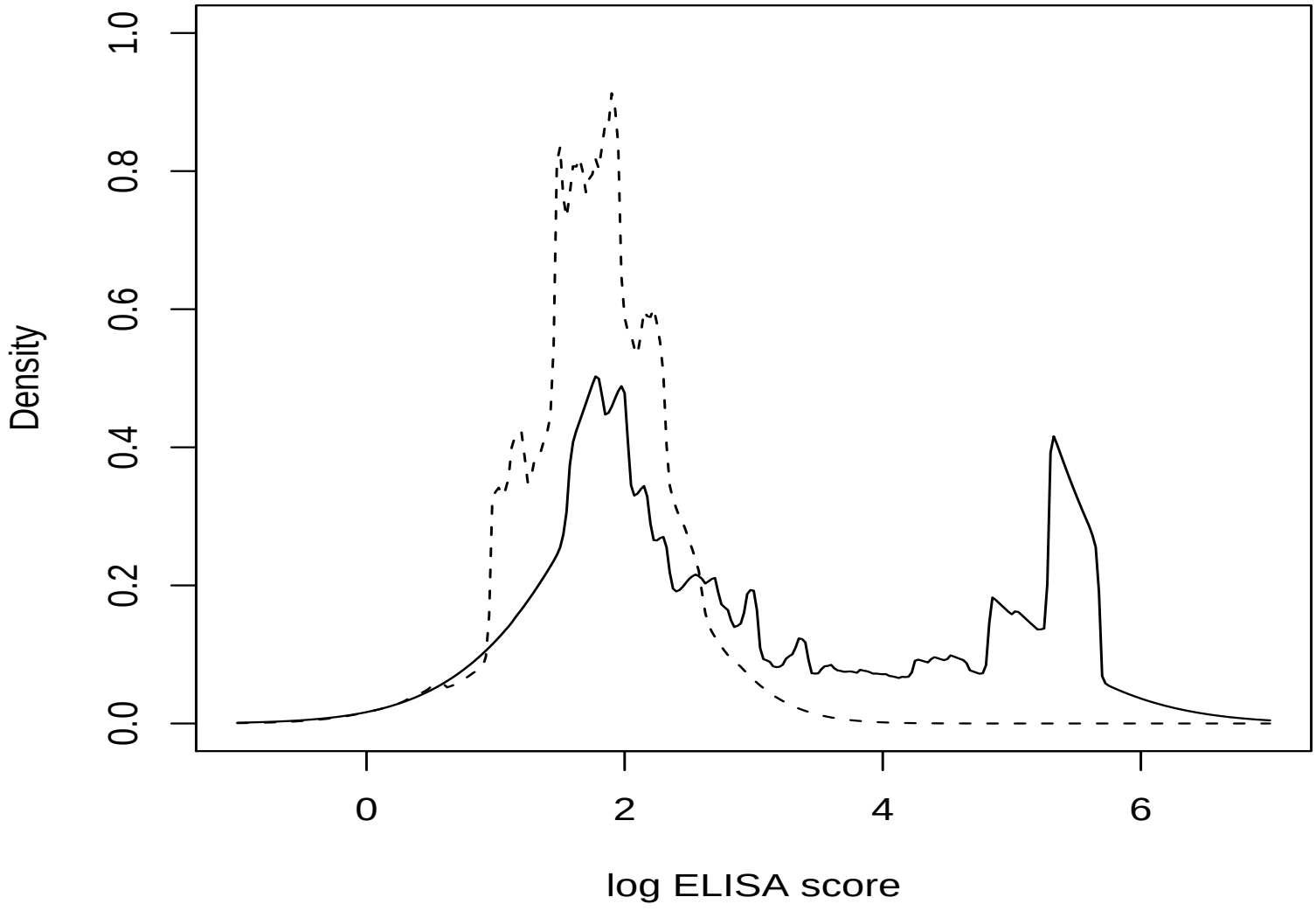


Example: Johne's Disease

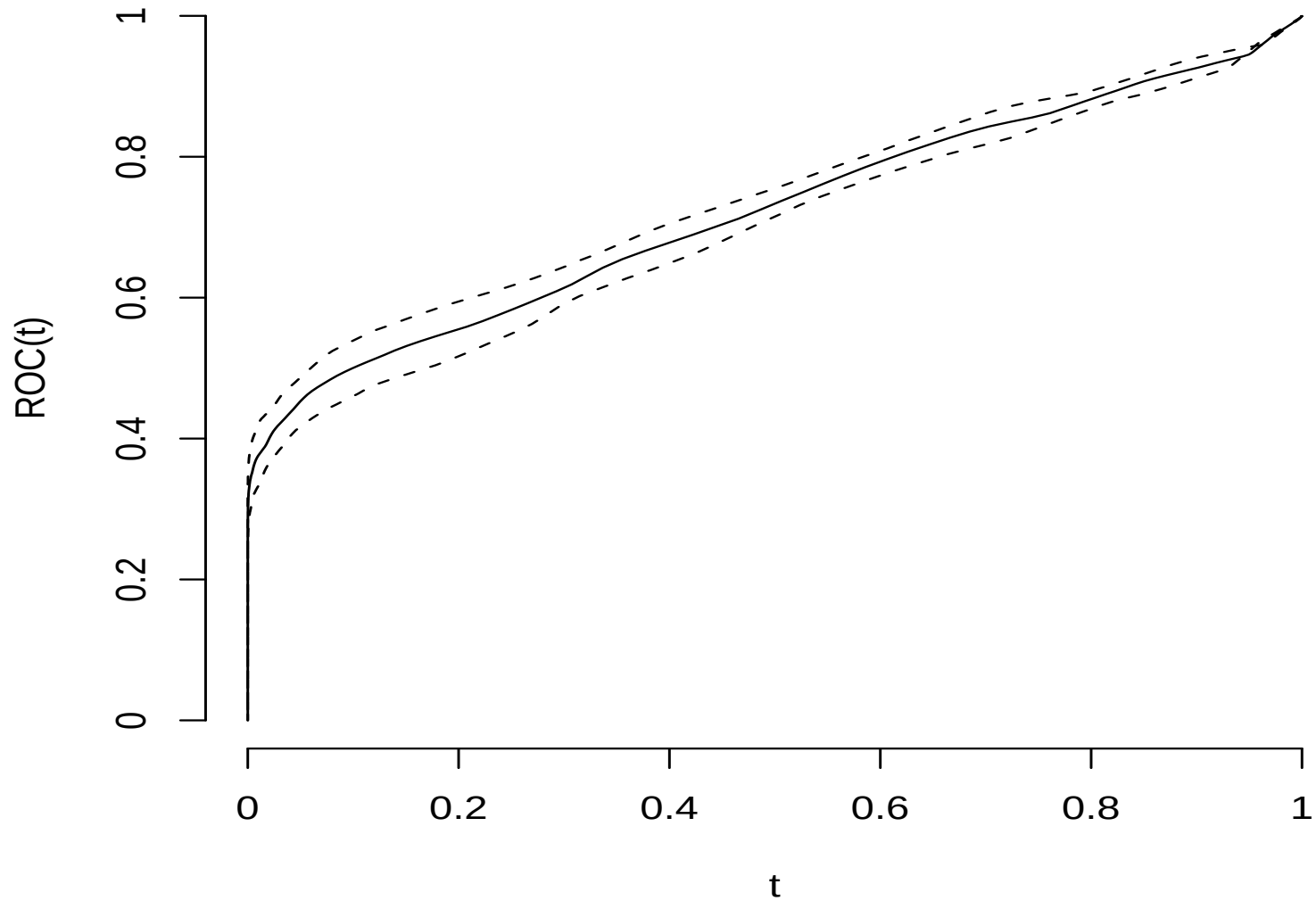
- ELISA scores used to screen for Johne's
- Sensitivity for standard cutoff is only around 0.30
- 599 cows tested
- We have an additional test based on culturing the virus (yes/no outcome); FC

- Specificity of FC is believed to be one
- Considered a range of beta priors on prevalence; inferences were insensitive to the choice
- Informative priors for baseline distributions based on expert opinion
- Estimated AUC = 0.71 (0.66, 0.76)

Johne's density estimates

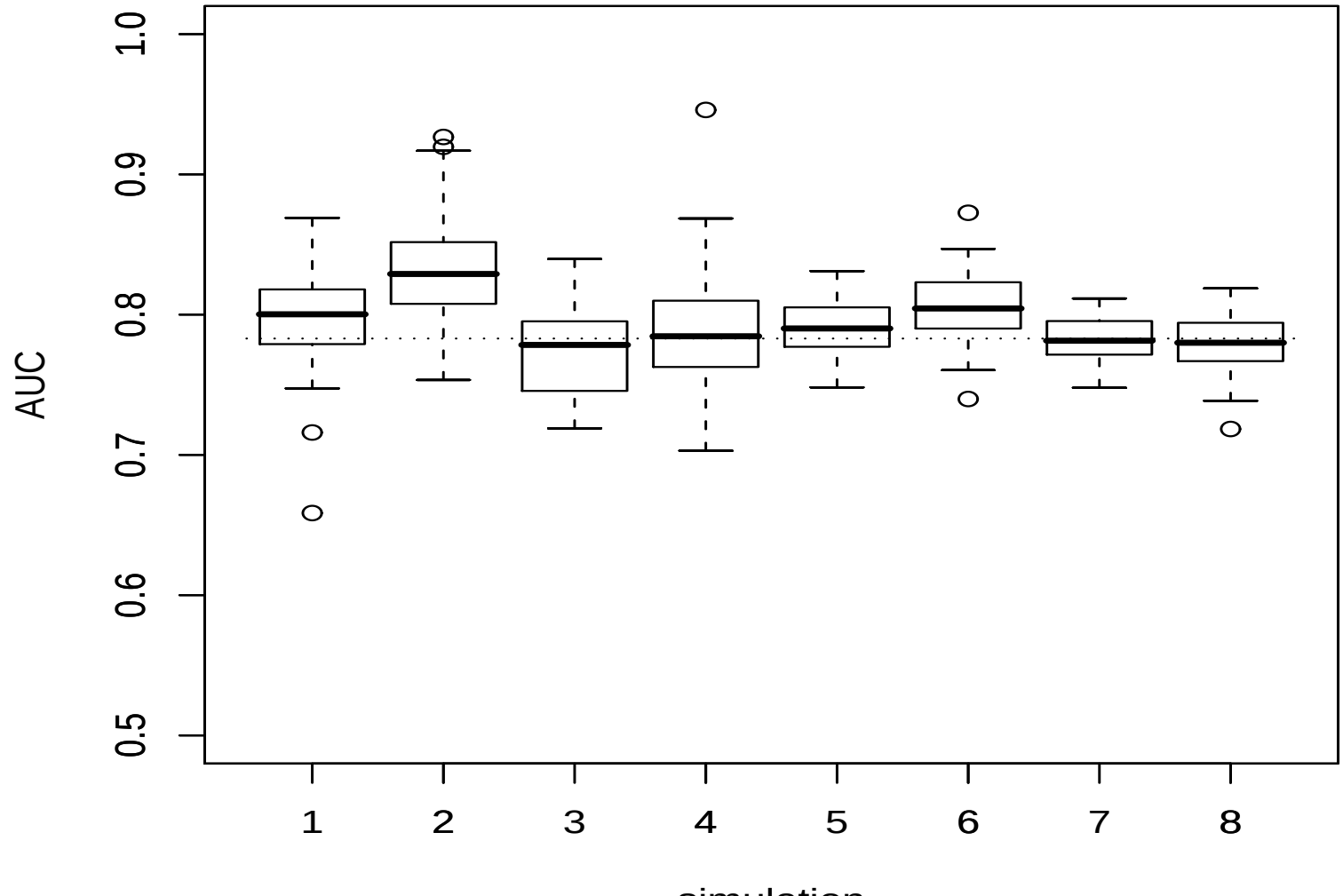


Johne's ROC curve estimates



Simulation: n=300, 1-4; n=1000, 5-8

Bin test Se=Sp=0.95: 1-2, 5-6; Se =0.85, Sp= 1: 3-4, 7-8



Semiparametric Models for Longitudinal Data: Application to Joint Modeling of Longitudinal Diagnostic Test Outcomes

Michelle Norris
Wes Johnson
Ian Gardner

Statistics and its Interface, 2009

Diagnostic Screening

- Test individuals repeatedly over time to detect infectoin
- Gold Standard test still unavailable
- Less expensive diagnostic outcomes are available eg serology, microscopy, polymerase chain reaction (PCR)
- Goals to estimate prevalence(s), diagnose individual subjects, and most importantly, to estimate Sensitivity as a function of time since infection
- Historically data have been cross-sectional
- Now consider models for longitudinal outcomes

The Data

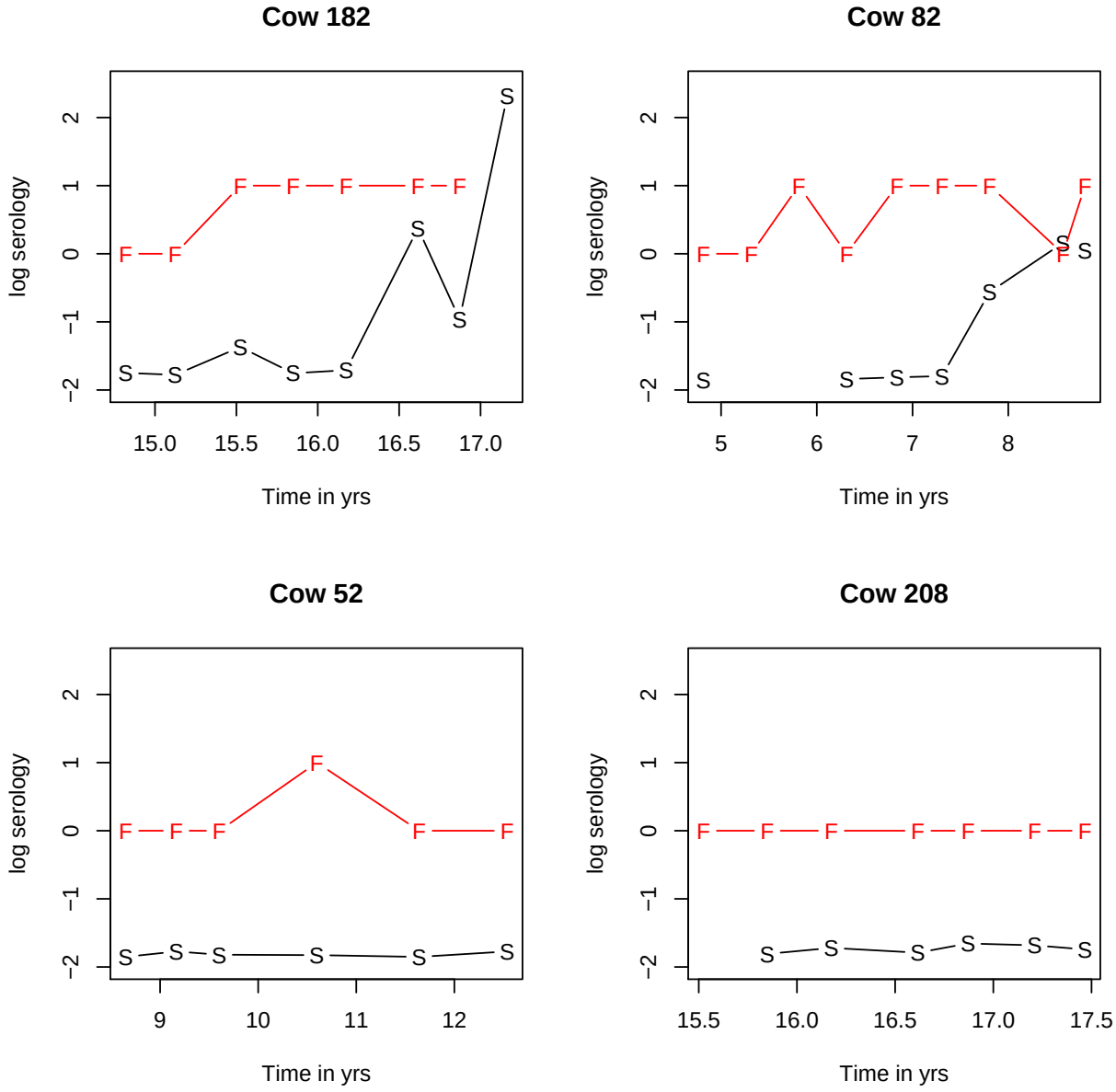
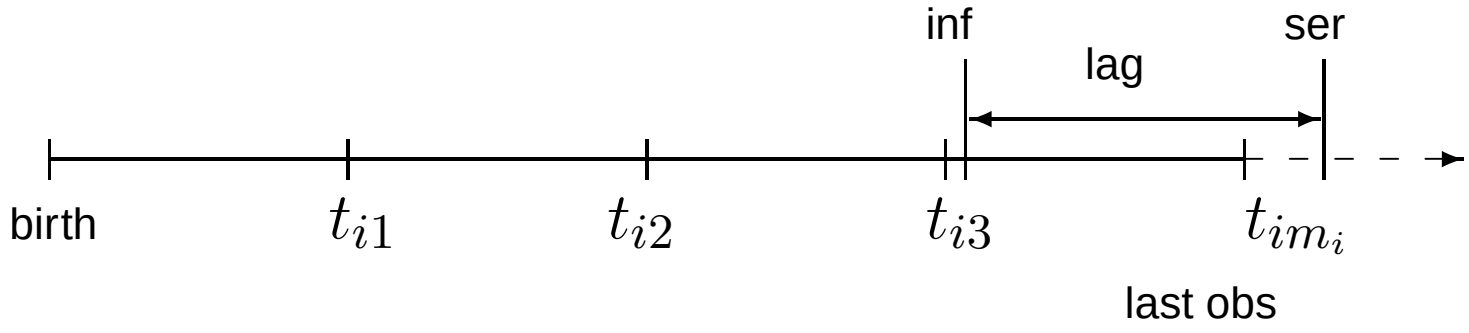
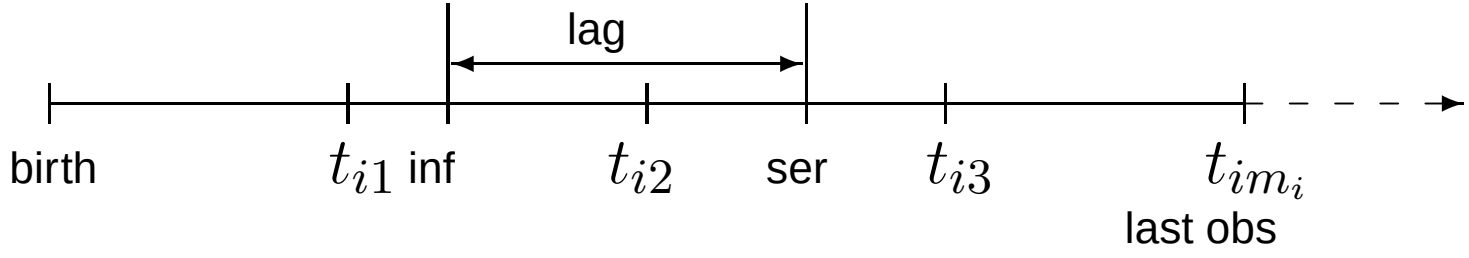

Cow 182
log serology
Time in yrs
Cow 82
log serology
Time in yrs
Cow 52
log serology
Time in yrs
Cow 208
log serology
Time in yrs

Figure: Latent Disease Status

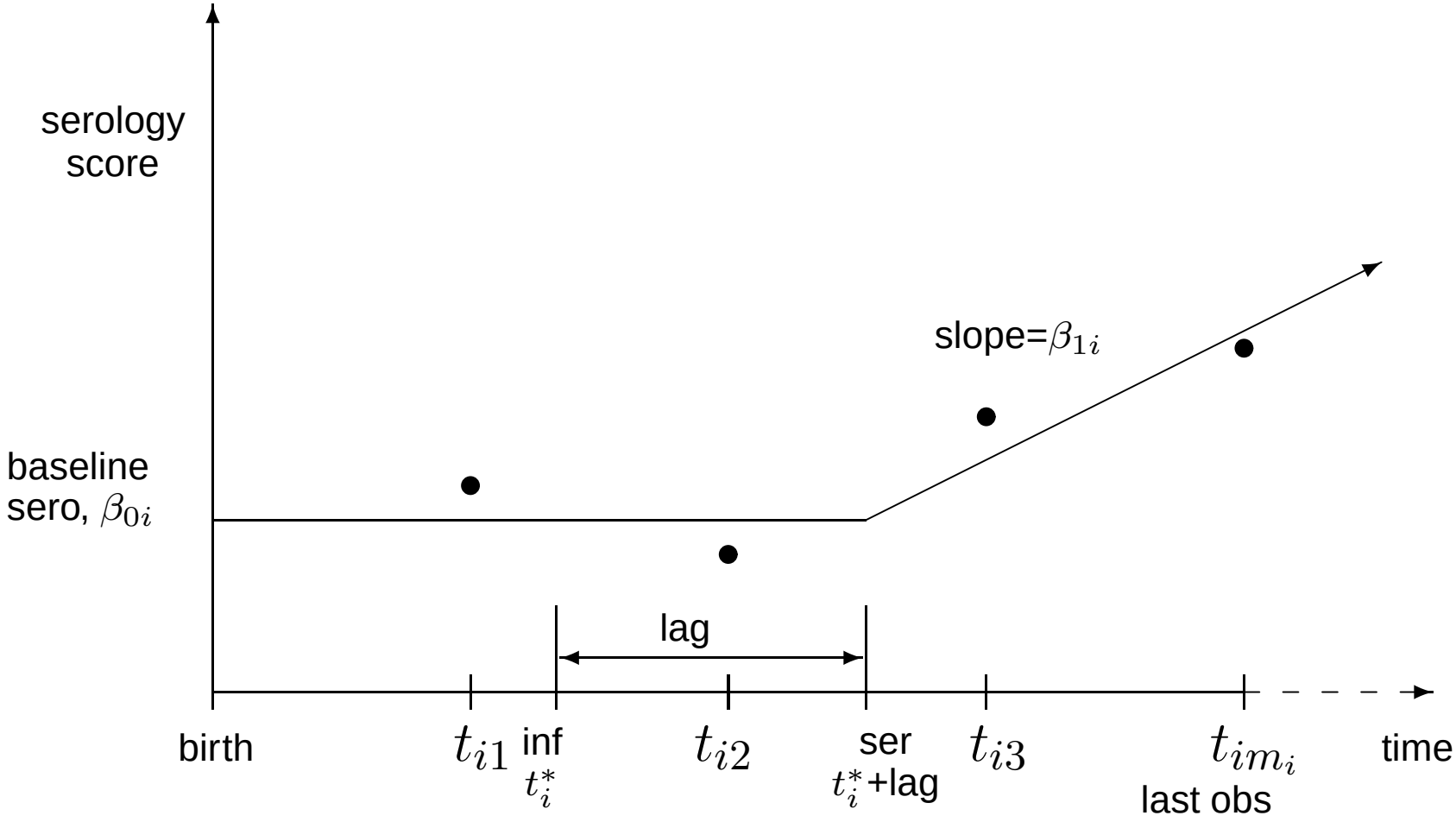


Infection, no Sero obs



Infection, Sero obs

Figure: Serology Trajectory



JD Data

- 365 cows, sampled approx every 6 months
- Number of obs per cow has median=6 , min=2 and max=23
- Serology scores log transformed for (approximate?) normality

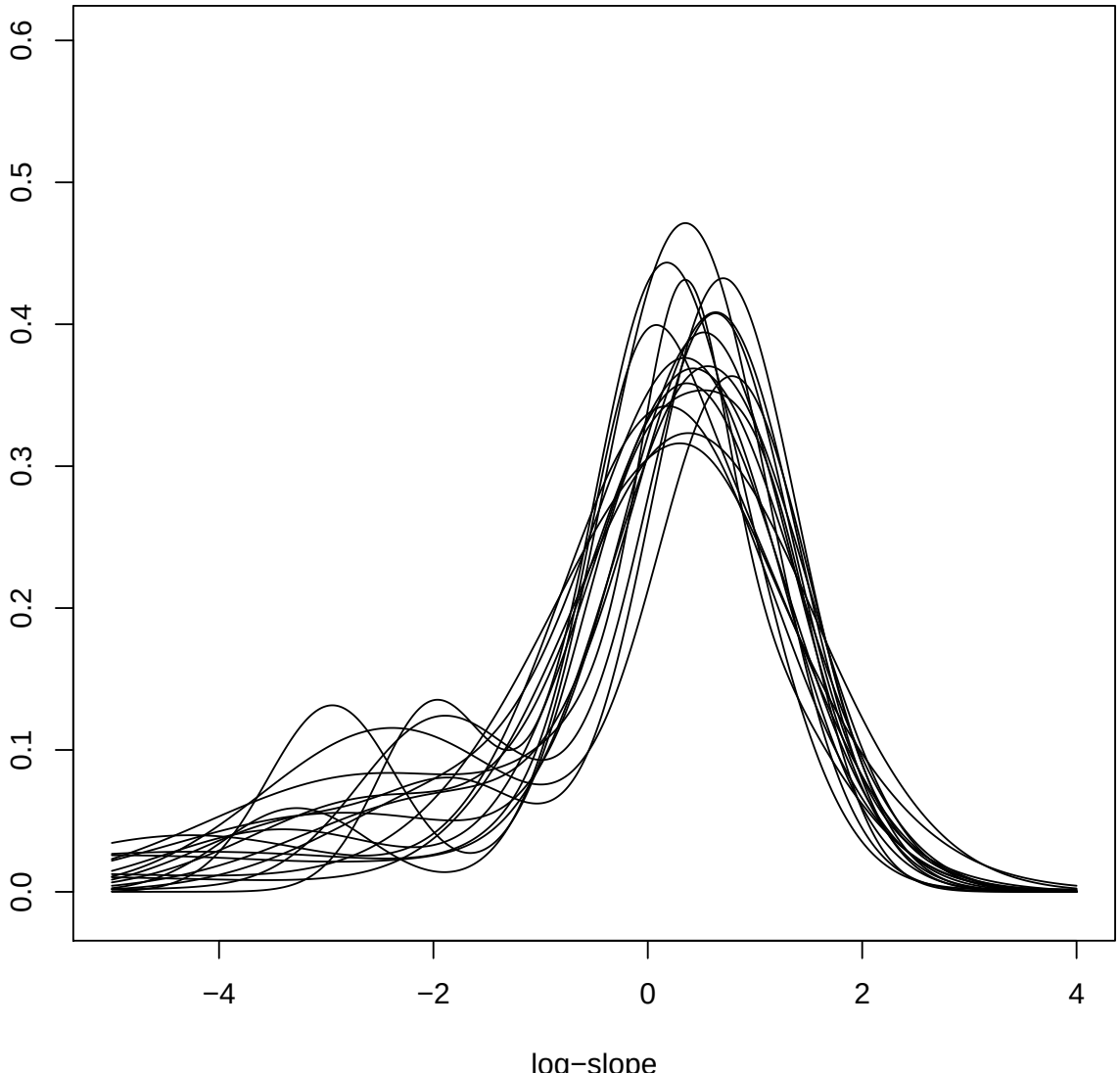
Semiparametric Model

- Parametric analysis used model with log-slopes as normally distributed; may be too restrictive
- The joint model specifies a DPM model for log-slopes:

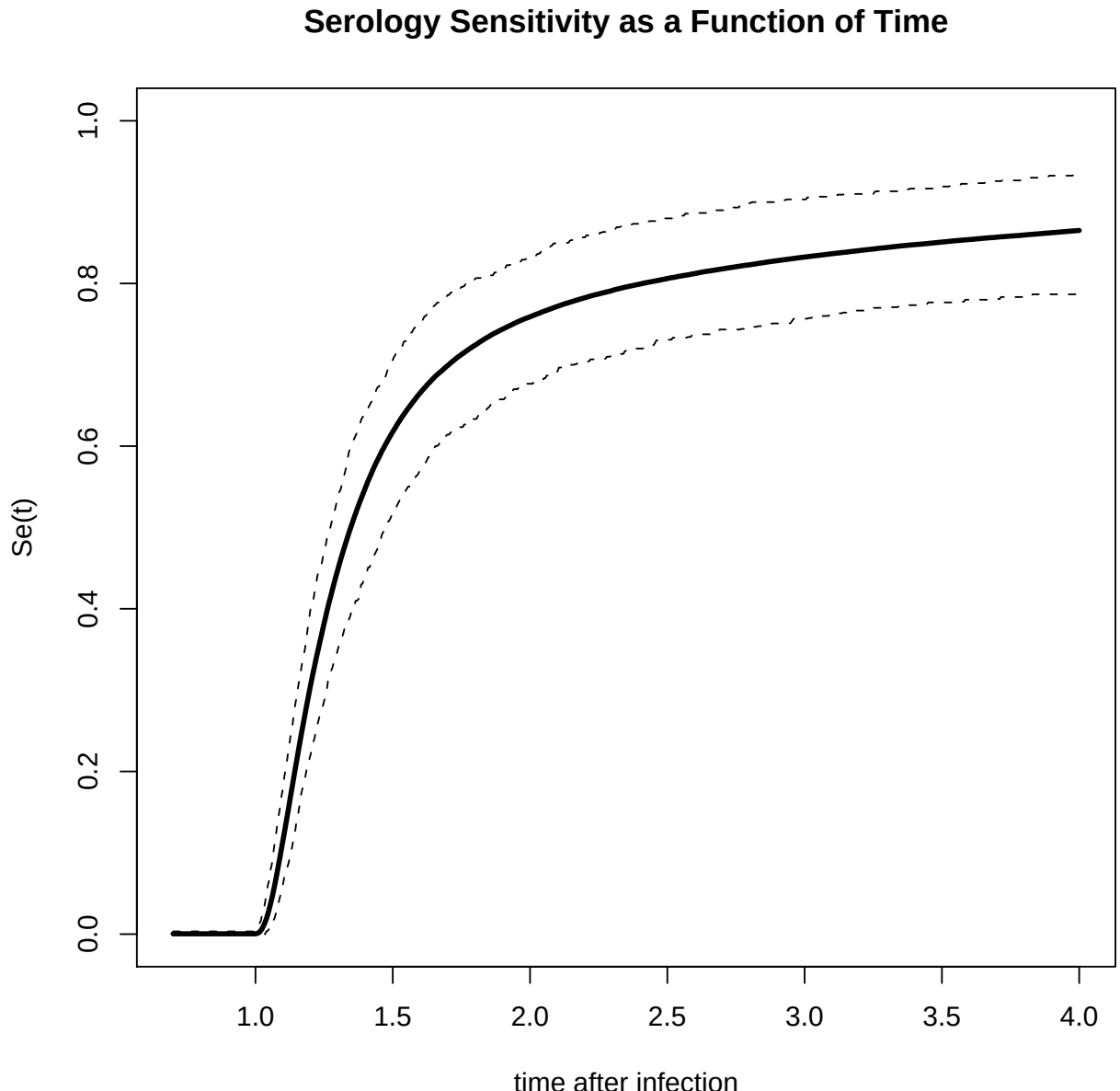
$$\begin{aligned} \log\beta_{1i} = \gamma_i \mid \mu_i, \tau_i &\overset{\perp}{\sim} N(\mu_i, \tau_i) \quad \text{for } i : k_i = 3 \\ (\mu_i, \tau_i) \mid G &\overset{\perp}{\sim} G \\ G \mid \alpha, G_0 &\sim DP(\alpha, G_0) \end{aligned}$$

Choose G_0 to be normal/inverse gamma conjugate: $\tau_i \sim \Gamma(\frac{s}{2}, \frac{S}{2})$ and $\mu_i \mid \tau_i \sim N(m, \frac{d}{\tau_i})$.

JD: Posterior Distn of slopes



JD: Serology Sensitivity



Bayesian Semiparametric Methods for Joint Modeling Longitudinal and Survival Data

Tim Hanson
Adam Branscum
Wes Johnson

Joint modeling setting

- Studies often involve an event/survival time of interest, and measurements on longitudinal processes that might be associated with patient prognosis

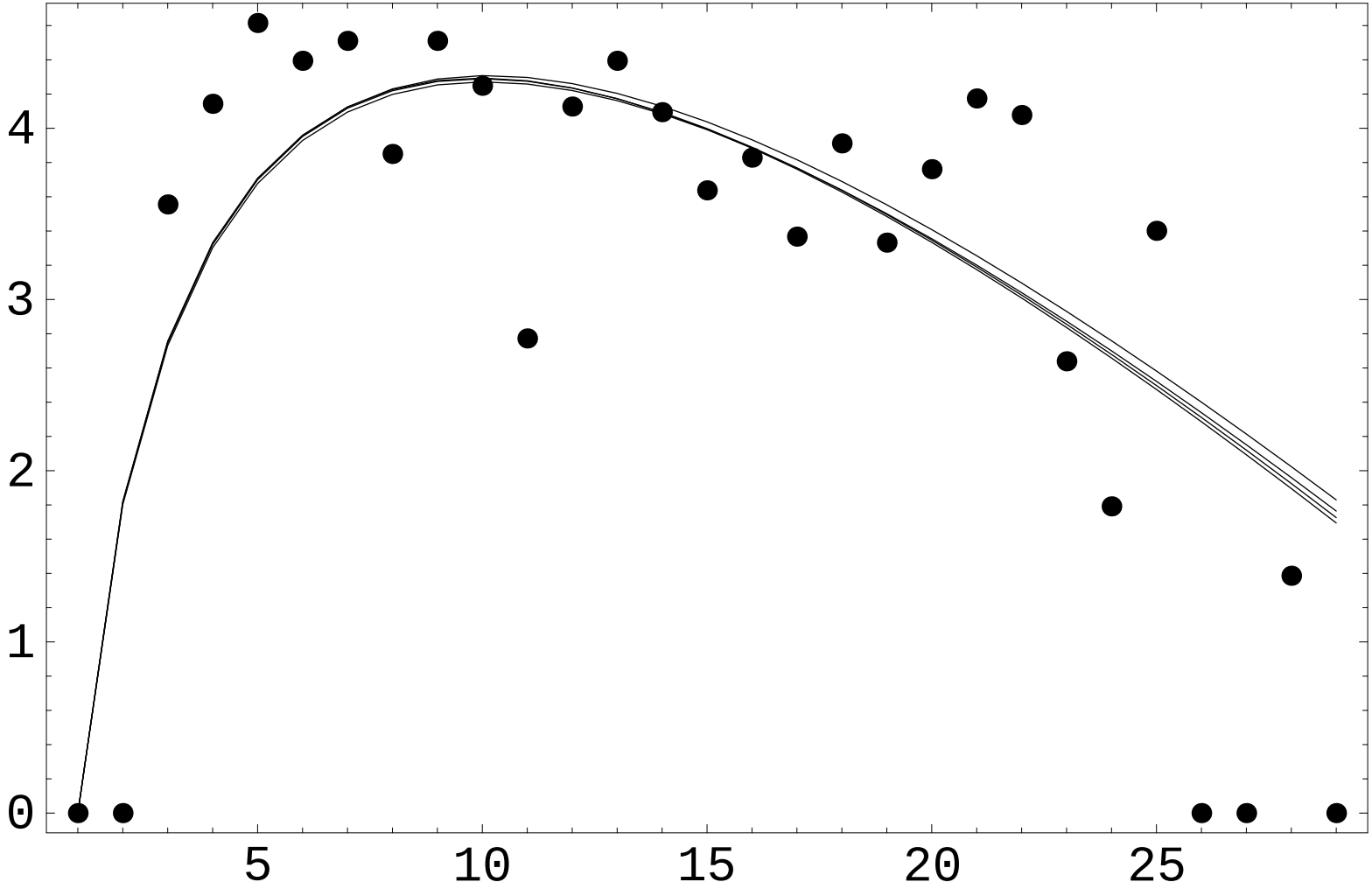
  1. Blood pressure measurements in dialysis patients $\rightarrow$ death

  2. Daily fertility counts in Mediteranean fruit flies $\rightarrow$ death

- Goals: Find trends in the time course of a longitudinal process and assess association between time-dependent processes and event prognosis

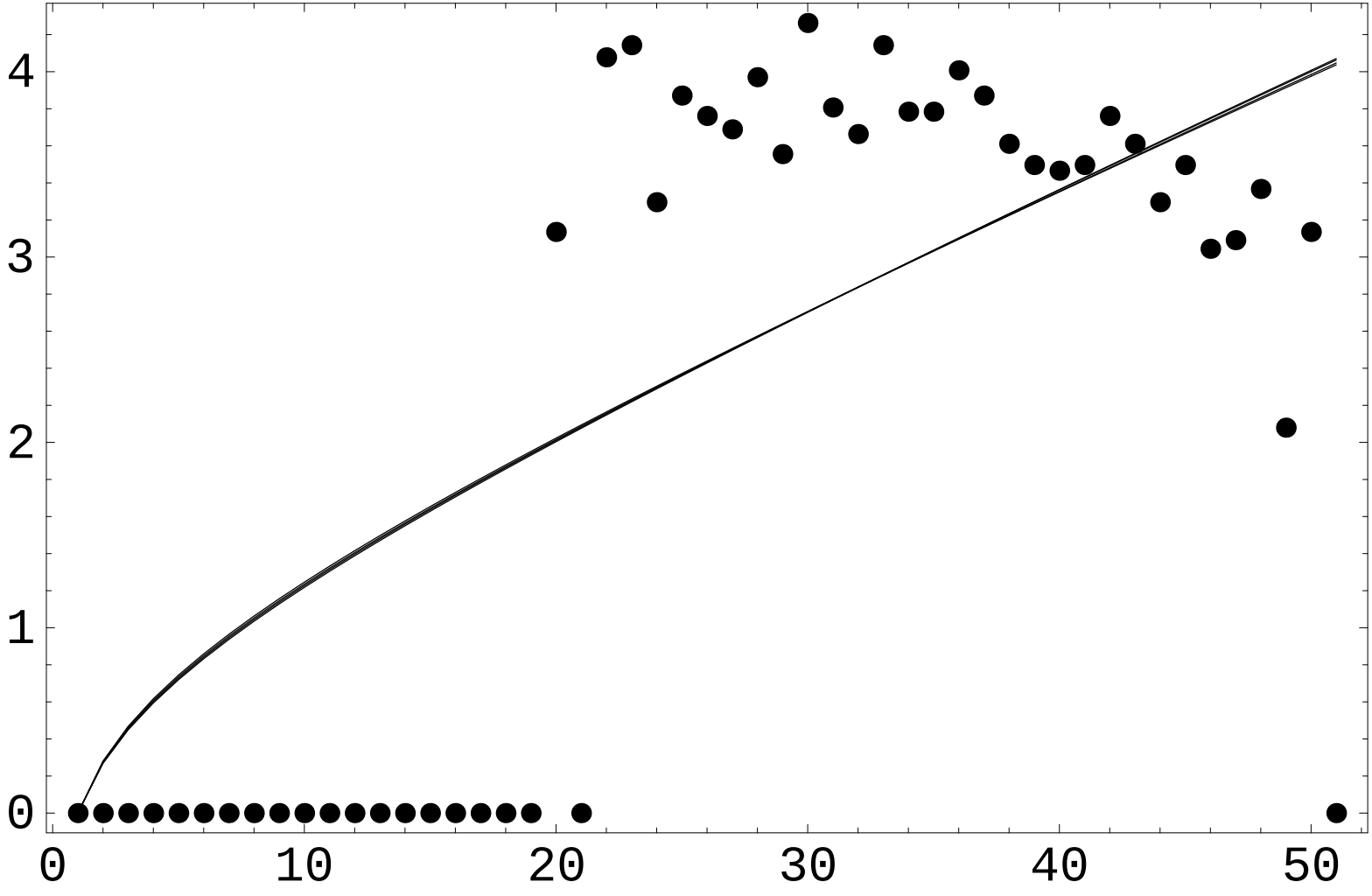
Fitted trajectory: Fly 1

- Fitted trajectory for a “typical" medfly. Similar shapes for PO, PH, CO, and longitudinal only analysis



Fitted trajectory: Fly 2

- Fitted trajectory for another medfly using PO, PH, CO, and longitudinal only analysis



Alternatives to Joint Modeling

- Ignore time dependent covariates (TDC)
- Standard survival analysis with TDCs, eg. treat longitudinal process as fixed
- Two-stage procedures, conducted by

  Modeling the observed (noisy) longitudinal process, and then

  Imputing the predicted process as if it were observed in a TDC survival model
- We compare joint analyses with alternative approaches

Longitudinal Component

- A common approach involves using models of the form:

$$\begin{aligned} y_i(t) &= x_i(t) + \epsilon_i(t) \\ x_i(t) &= \mathbf{f}(t)\boldsymbol{\gamma} + \mathbf{g}(t)\mathbf{b}_i + U_i(t) + z_i\boldsymbol{\alpha} \\ \epsilon_i(t) &\overset{iid}{\sim} N(0, \sigma^2) \end{aligned}$$

- $\mathbf{f}(t)$ and $\mathbf{g}(t)$ are vectors of known functions of time or basis functions eg. splines, wavelets
- γ is fixed; $\mathbf{b}_i$'s are random
- $U_i(t)$ is a mean-zero stochastic process, eg. an Ornstein-Uhlenback process

Survival Models: “Raw"

Let $Y_t = \{y(s) : s < t\}$ denote the raw history

- Previous approaches: Cox (1972)

$$h(t|Y_t) = e^{y(t)\beta} h_0(t)$$

- Prentice (1982) and Hanson et al (2009)

$$h(t|Y_t) = e^{y(t)\beta} h_0(te^{y(t)\beta})$$

- Cox and Oakes (1984)

$$h(t|Y_t) = e^{y(t)\beta} h_0(\bar{c}(t)t), \ \bar{c}(t) = \frac{1}{t}\int_0^t e^{y(s)\beta} ds$$

- Sundaram (2006) extended the Proportional Odds model to TDCs

Survival Models: Imputation

- Classic assumption is that $y(t)$ is constant between observation times; problematic if time points are irregularly spaced with long periods between them
- May lead to biased estimates of regression coefficients
- Modeling the longitudinal process allows for incorporation of measurement error, and also allows for imputation of values of the process between observation times
- Survival modeled as a function of predicted “true" history, $\hat{X}_t = \{\hat{x}(s) : s \leq t\}$

Bayesian Semi-parametric Models

- Parametric approaches specify *baseline* S_0 as, eg., log-logistic, normal...

- Hanson and Johnson (2002) modeled S_0 with a MFPT:

$$S_0 \sim \int PT(c, G_\psi) p(d\psi, dc)$$

c is weight parameter

G_ψ is a standard parametric distribution (eg. log-logistic)

Model choice

- We compare PO, Cox and PO models using predictive method of Geisser and Eddy (1979)
- Cross-validatory (pseudo marginal likelihood (PML)) criterion:

$$PML = \prod_i p(T_i|\mathbf{T}_{-i}, y_{1:n})$$

- Ratios of these for distinct models mimic Bayes factors
- Define

$$LPML = \sum_{i=1}^{n} \ln(PML_i)$$

- Find model with maximum LPML

Illustration: Medfly Data

- Data from a study of reproductive patterns of female Mediterranean fruit flies. Recorded the number of eggs produced each day throughout lifespans
- Goal was to examine the association between egg production patterns and lifetime
- A frequentist approach was used by Tseng et al (2005)
- Sample size of 251 flies with lifespans ranging from 22 to 99 days, and no censored observations

Illustration: Medfly Data

- Longitudinal structure for egg production that Tseng et al. used was based on $\ln(y_i(t)+1)$ with

$$x_i(t|\mathbf{b}_i) = b_{1i}\ln(t) + b_{2i}(t-1)$$

- Longitudinal correlation structure is not modeled.
- For (raw) analyses that treated egg production as a fixed TDC, we used a piecewise constant function
- Used methods of Hanson, Johnson, and Laud (2009)

Model comparison

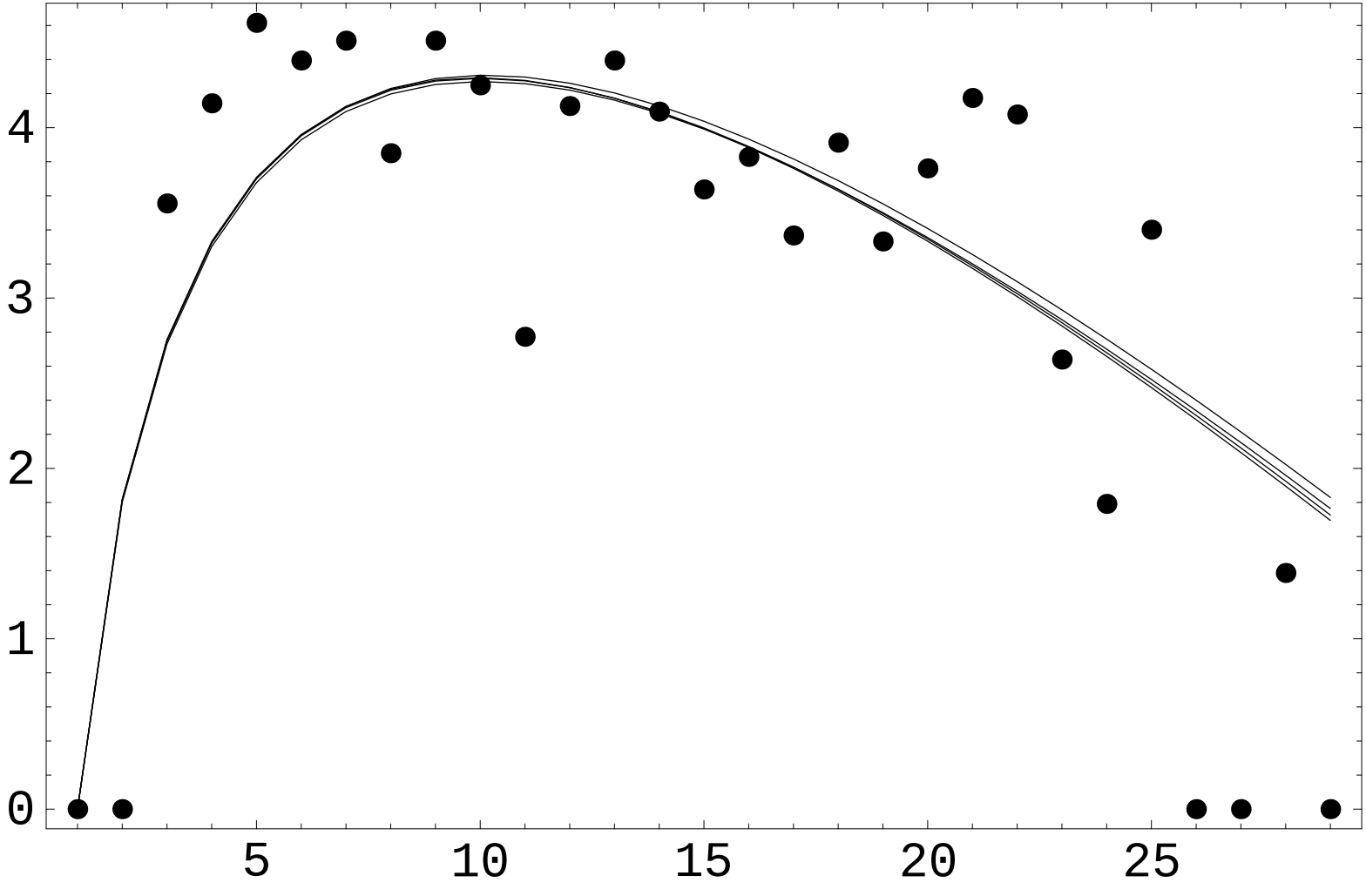
- *negative*-LPML statistics (smaller is better) comparing modeling approaches:

Model	Method	PO	PH	CO
parametric	raw	867	870	937
MPT	raw	865	866	938
MPT	imputed	947	959	973
parametric	joint	947	959	973
MPT	joint	945	956	973

- exp(difference in LPML) = pseudo Bayes factor

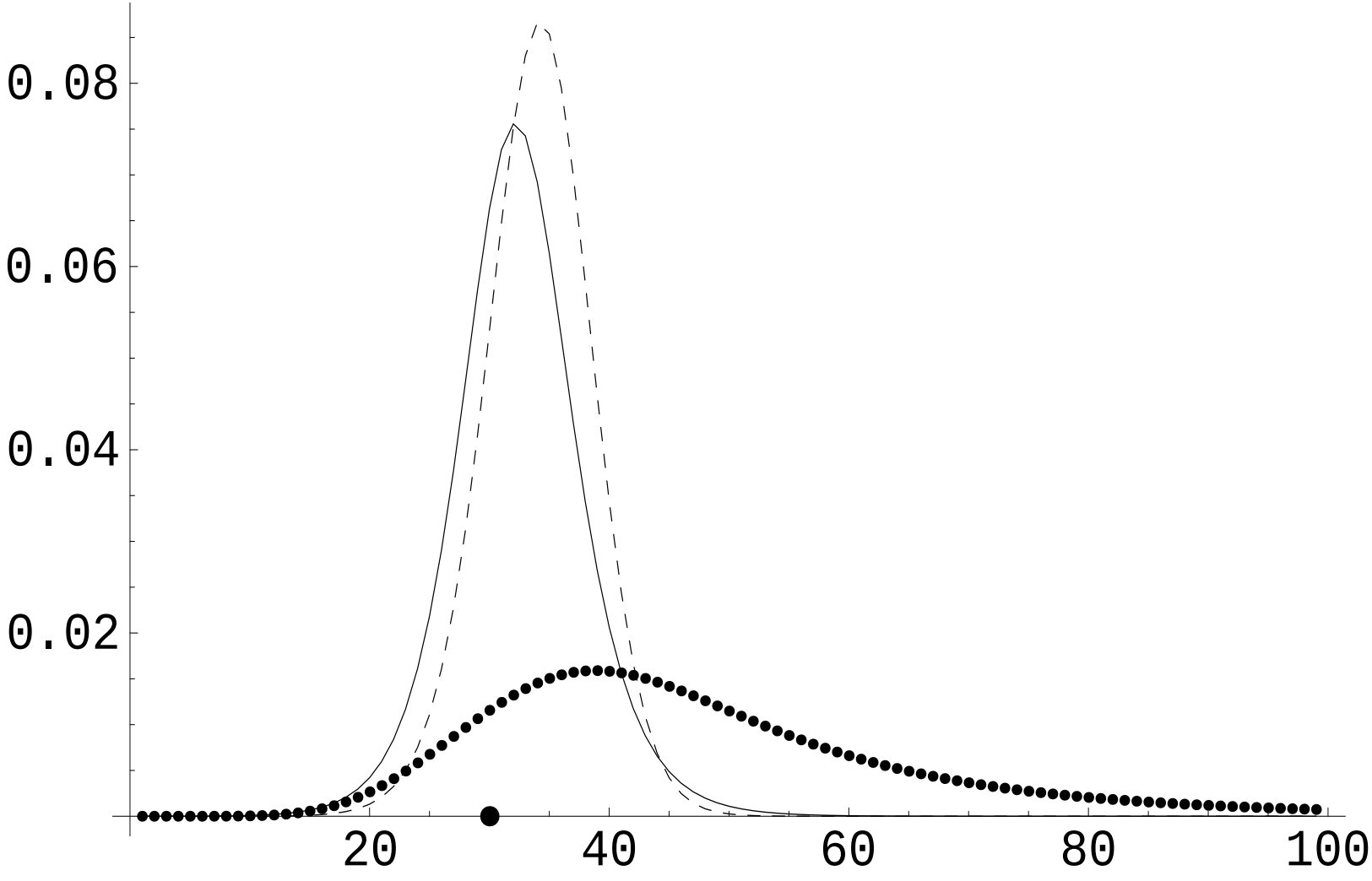
Fitted trajectory: Fly 1

- Fitted trajectory for a “typical" medfly. Similar shapes for PO, PH, CO, and longitudinal only analysis



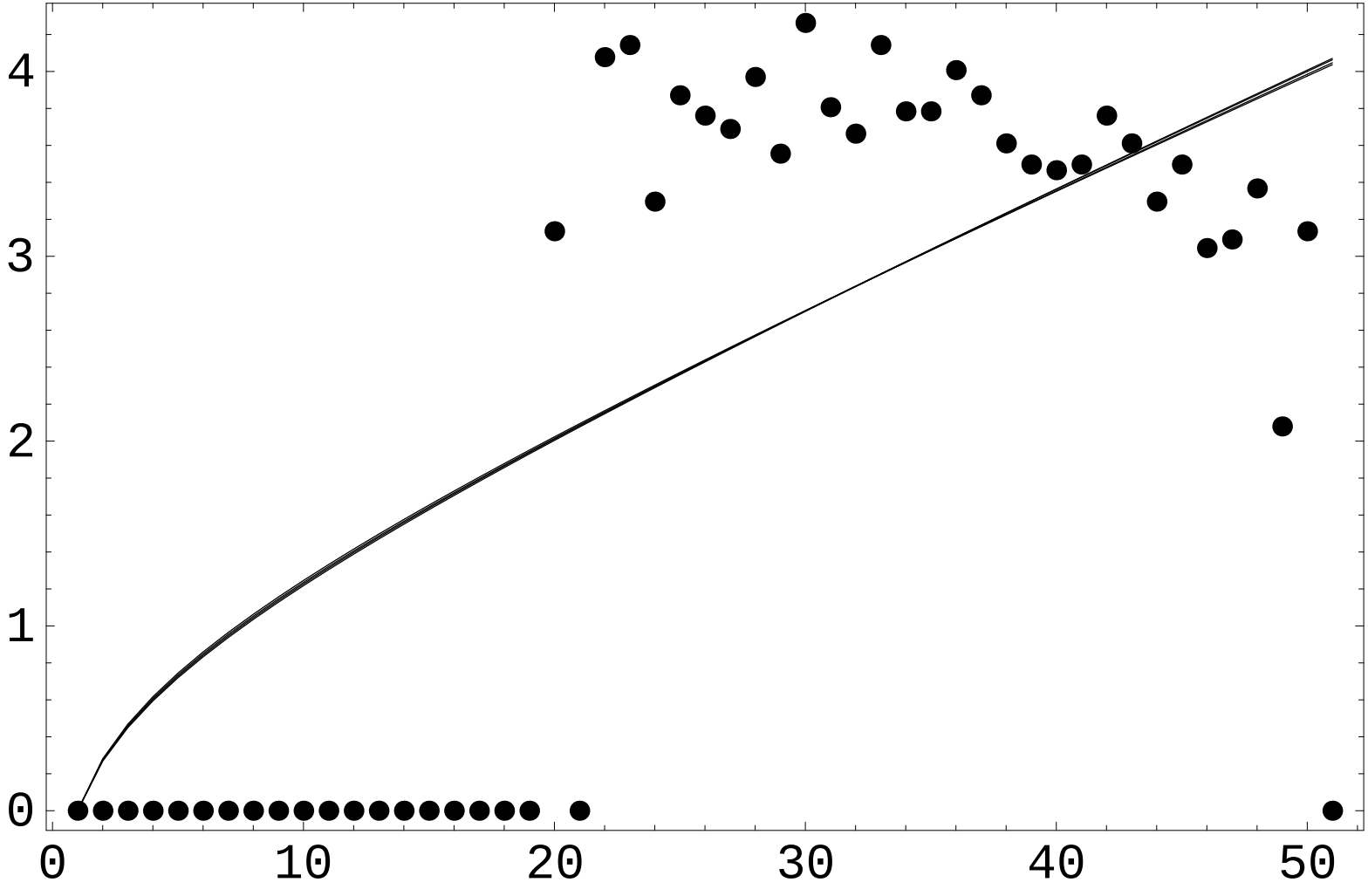
Predictive survival density: Fly 1

- Solid is PO, dashed is PH, and dotted is CO



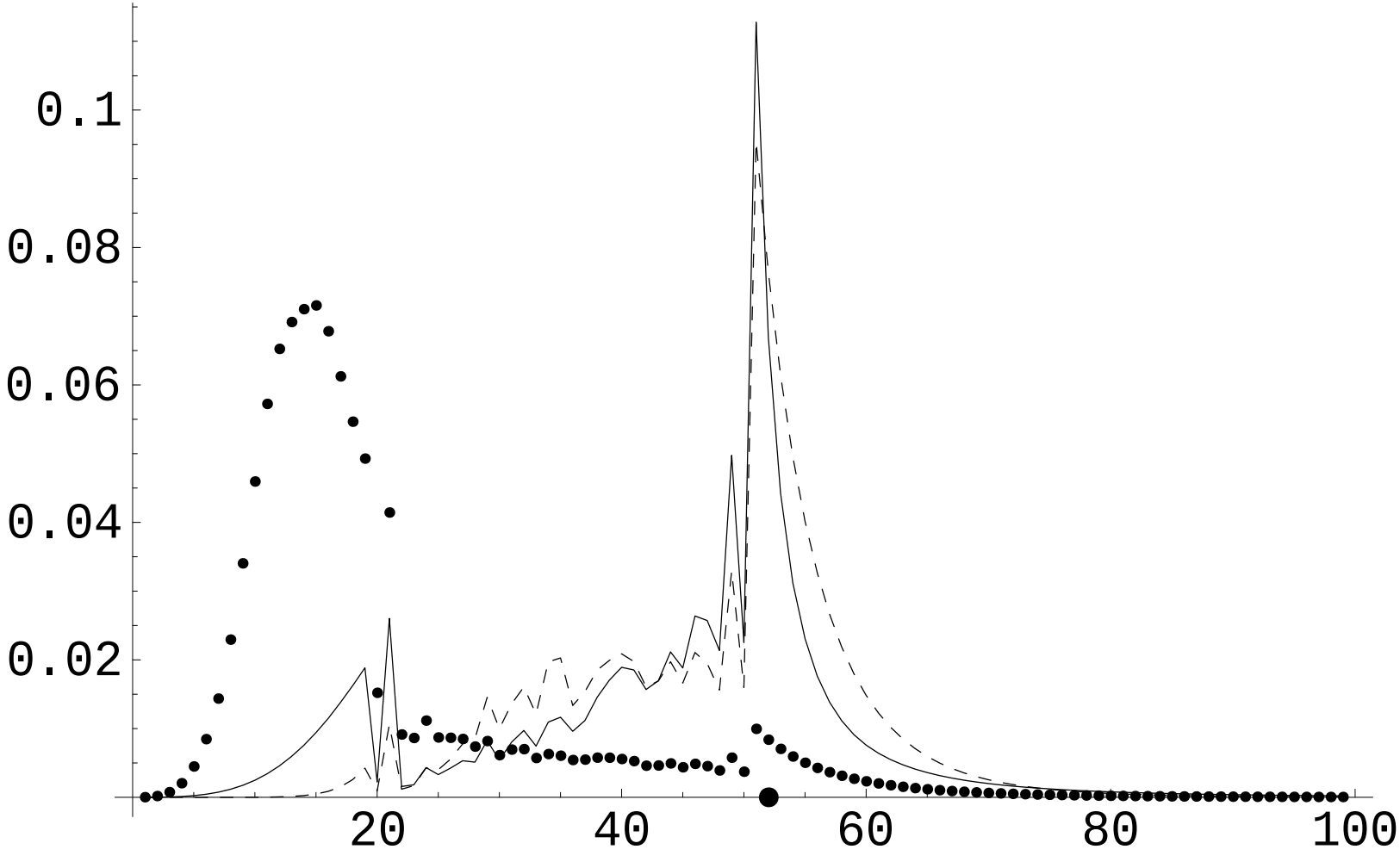
Fitted trajectory: Fly 2

- Fitted trajectory for another medfly using PO, PH, CO, and longitudinal only analysis



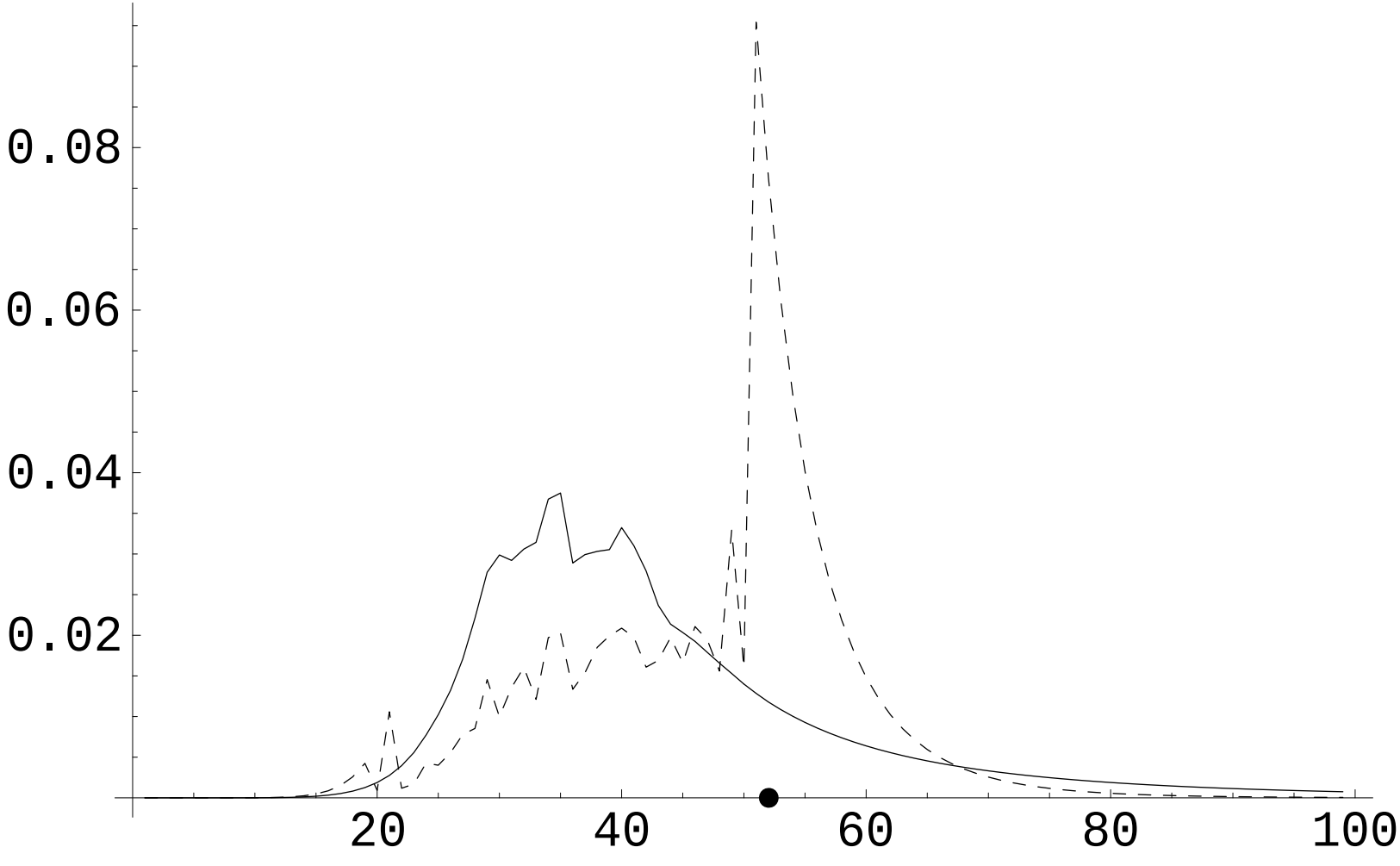
Predictive survival density: Fly 2

- Semiparametric PO (solid), PH (dashed) and CO (dotted) analyses using **raw** trajectories.



Predictive survival density: Fly 2

- Semiparametric PH analyses comparing raw trajectories (dashed line) to **joint** analysis (solid line).



Conclusion

- We replaced $x(t)$ with a penalized spline, but raw analysis was still preferable
- Joint modeling appears to not be necessary for these data

Bayesian Nonparametric, Non-Proportional Hazards Survival Analysis

Maria De Iorio
Wes Johnson
Peter Müller
Gary L. Rosner

Biometrics, 2009

Survival Analysis

- Use Dependent Dirichlet Process (MacEachern, 1999)
- Illustration based on a cancer clinical trial
- Survival probabilities for early times are est lower for high dose treatment than for low dose
- The reverse is true later for later times, possibly due to toxic effect of the high dose for those less healthy at beginning of study

DDP Regression

- Model: $T|\, x,\, \beta,\, \tau \sim LN(x\beta, 1/\tau)$
- $(\beta, \tau)|\, G \sim G, \quad G \sim DP(c, F_0)$
- Then the linear DDP model can be written as

$$\begin{aligned} f(t \mid x, G) &= \int f(t \mid x\beta, \tau) dG(\beta, \tau) \\ G &\sim DP(c, F_0) \end{aligned}$$

- Let G_x denote the random CDF for $T|x$

Equivalence

- Suppose $G_x \sim DP(c, F_x)$ for all $x \in \mathcal{X}$, eg.

$$G_x = \sum_{h=1}^{\infty} p_h \delta_{\lambda_{xh}}(\cdot), \text{ all } x \in \mathcal{X}$$

- Model $\lambda_h = \{\lambda_{xh},\ x \in \mathcal{X}\} \overset{iid}{\sim} p(\lambda)$
- Define F_x to be the CDF corresponding to $x\beta, \quad \beta \sim F_0$
- This induces $p(\lambda)$
- Then G_x and G_{x*} are dependent by virtue of the modeled relationship between the random pairs $\{(\lambda_{xh}, \lambda_{x*h}) : h = 1, 2, ...\}$
- The resulting collection $\{G_x : x \in \mathcal{X}\}$ is said to have a DDP distribution (MacEachern, 1999)

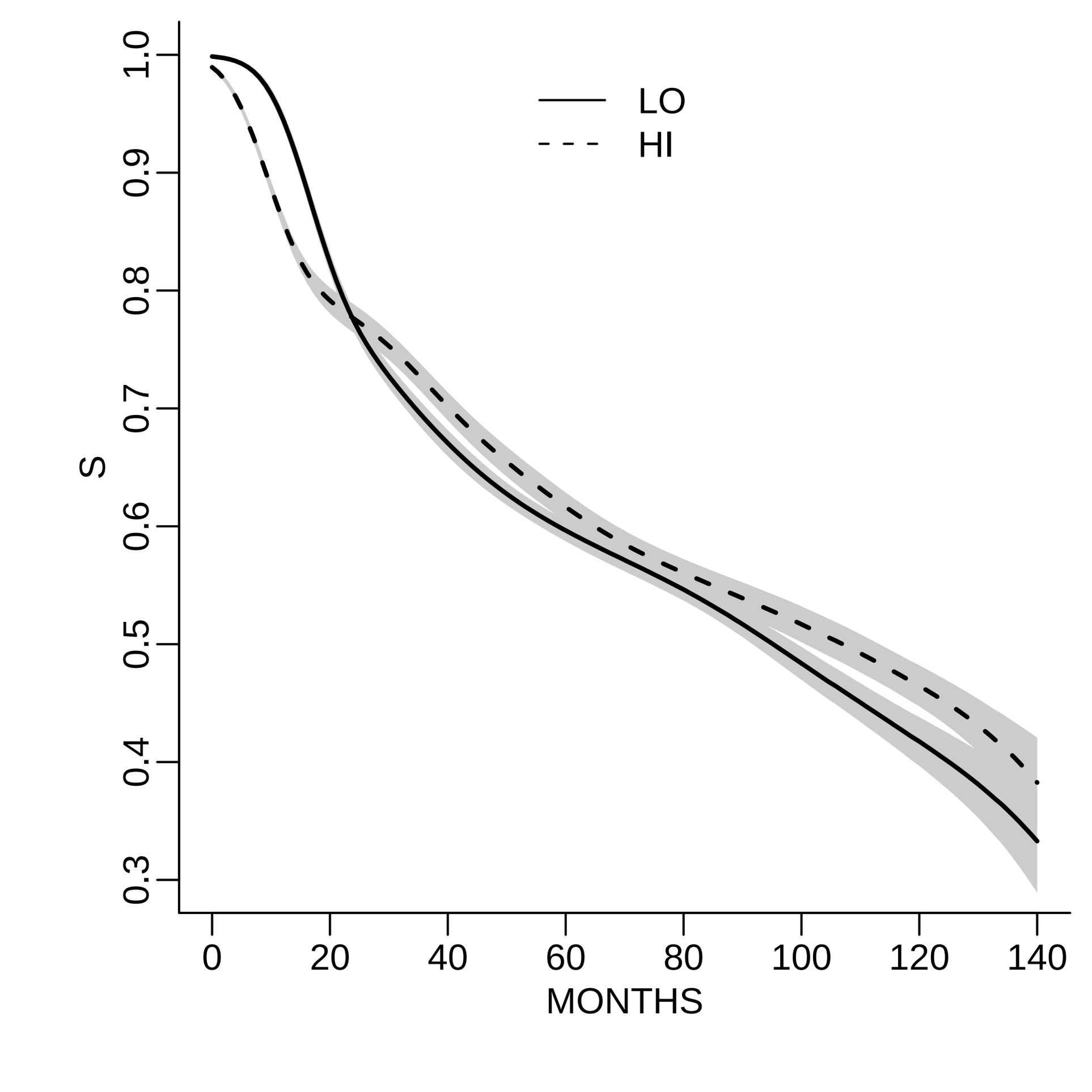
Cancer Clinical Trial

- The data record the event-free survival time in months for 761 women
- 53% censoring
- Determine if high doses of the treatment are more effective than lower doses
- High doses of treatment are known to be associated with a high risk of treatment related mortality

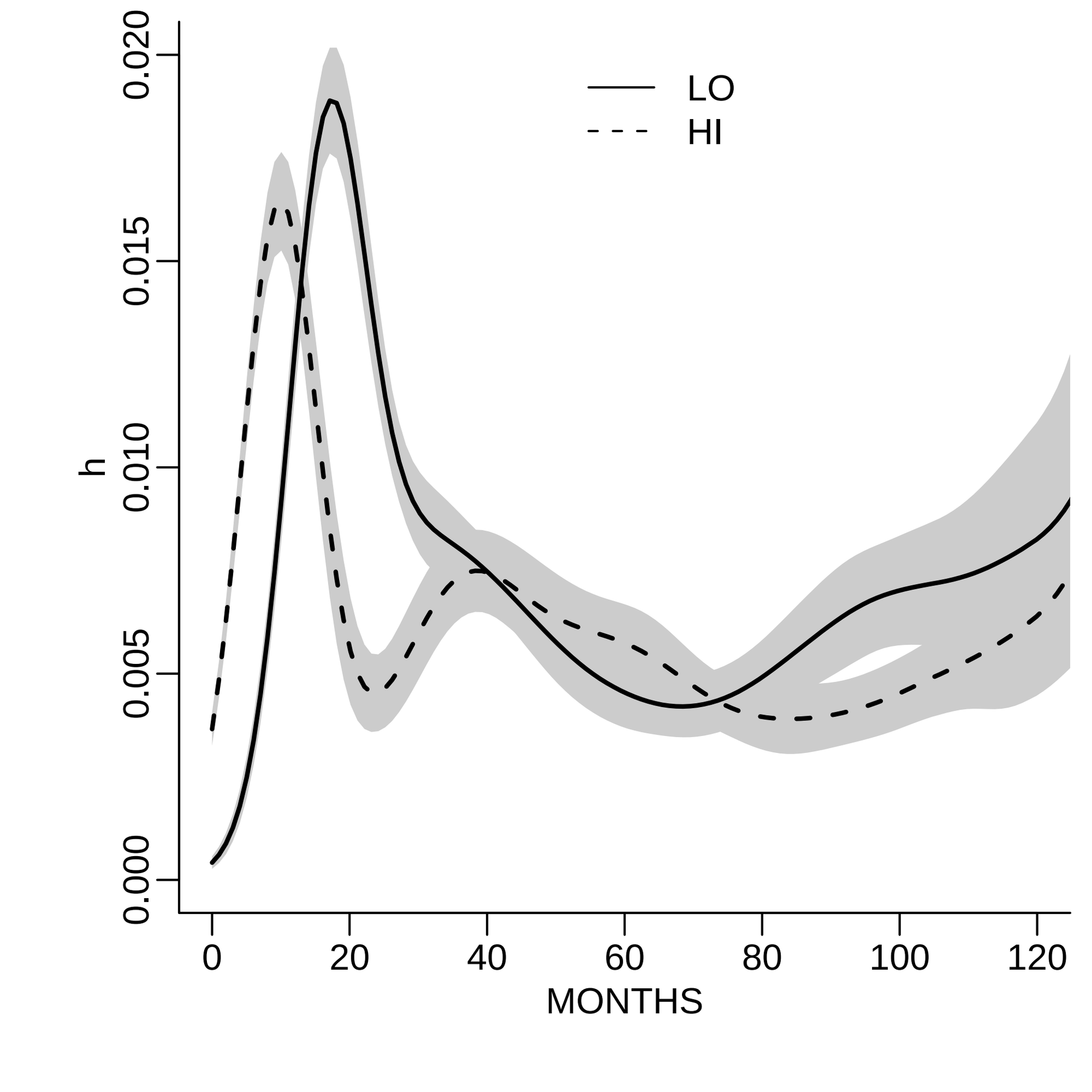
Cancer Clinical Trial

- Clinicians hope initial risk offset by reduction in mortality; justifying more aggressive therapy
- Three categorical covariates plus an interaction:
  - Treatment dose (low/high)
  - Estrogen receptor (ER) status (pos/neg)
  - Tumor Size (TS)
  - Dose by ER interaction

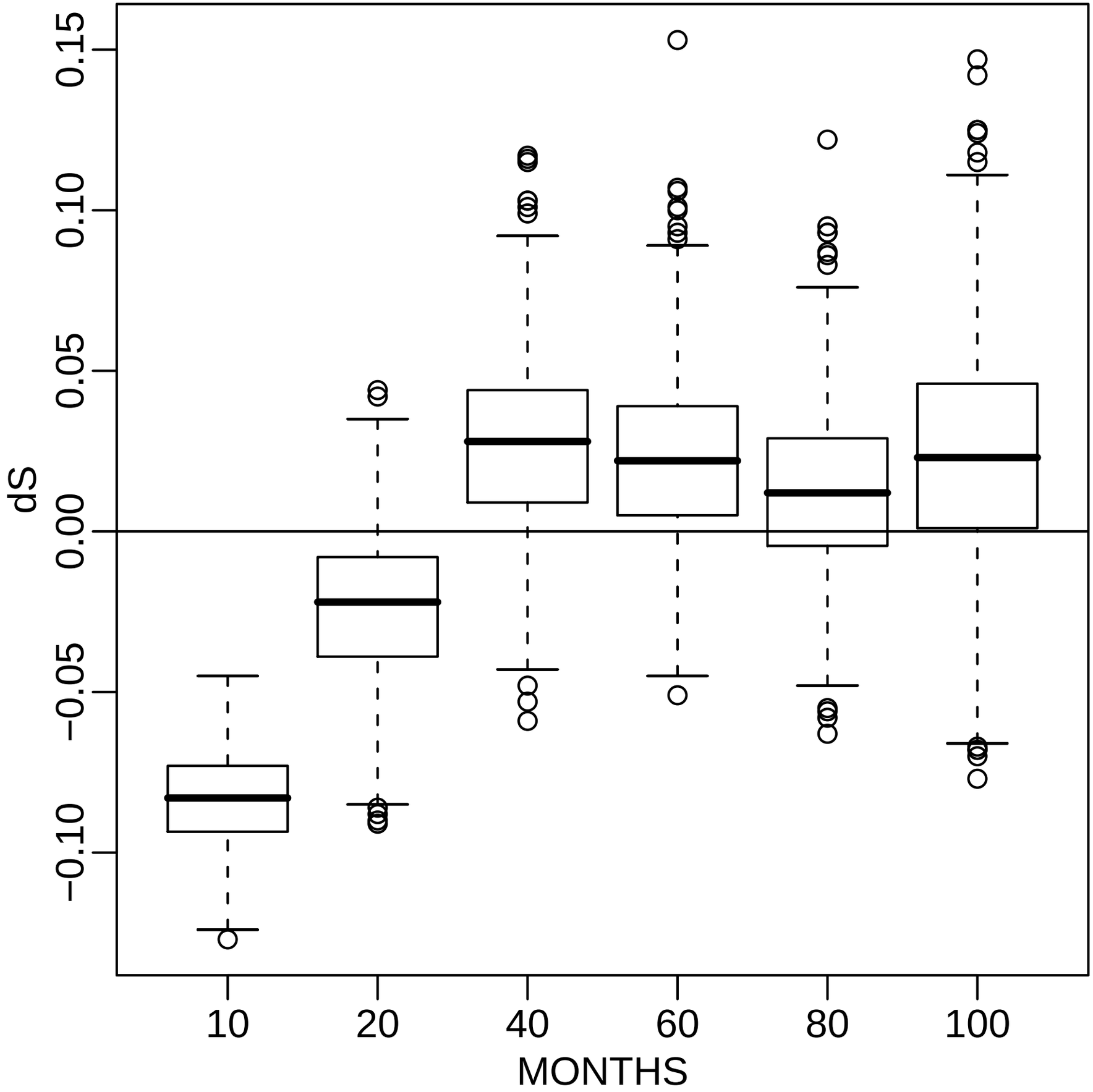
Post Surv: L vs H; Small/ER+


LO
HI
S
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
20
40
60
80
100
120
140
MONTHS

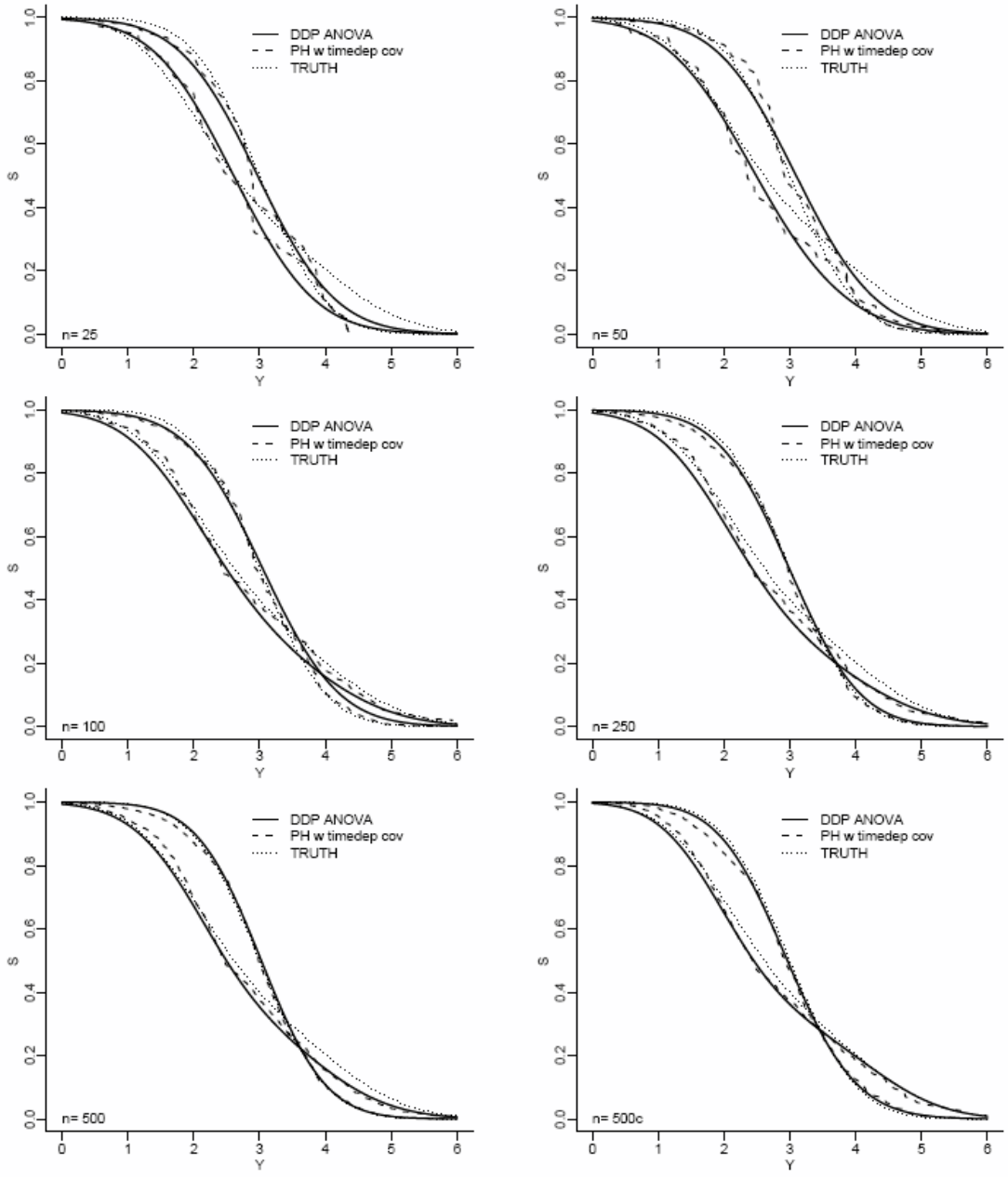
Hazards: L vs H: Small/ER+


LO
HI
h
0.000
0.005
0.010
0.015
0.020
0
20
40
60
80
100
120
MONTHS

Posterior of diff in Surv


dS
0.15
0.10
0.05
0.00
−0.05
−0.10
10
20
40
60
80
100
MONTHS

Simulation


DDP ANOVA
PH w timedep cov
TRUTH
n= 25
n= 50
n= 100
n= 250
n= 500
n= 500c
S
Y